Representation and Reasoning

A Causal Model Approach

By Milena Nikolic

Cognitive, Perceptual and Brain Sciences Research Department

University College London

Thesis submitted for the degree of

Doctor in Philosophy (PhD)

December, 2013

Thesis Declaration

I declare that this thesis was composed by myself and that the work contained herein is my own except where explicitly stated otherwise in the text. This work has not been submitted for any other degree or professional qualification.

Signature:

London, 10 December 2013

Abstract

How do we represent our world and how do we use these representations to reason about it? The three studies reported in this thesis explored different aspects of the answer to this question. Even though these investigations offered diverse angles, they all originated from the same psychological theory of representation and reasoning. This is the idea that people represent the world and reason about it by constructing dynamic qualitative causal networks. The first study investigated how mock jurors represent criminal evidence and reason with such representations. The second study examined how people represent the causes of a complex environmental problem and how their individual representations are directly linked to how they reason about the issue. The third and final study inspected how people represent causal loops and reason in accordance with these cyclical representations. These studies suggest that people do represent the world by arranging evidence, causes, or pieces of information into a causal network. In addition, the studies support the idea that these networks are of a qualitative nature. All three studies also indicated that people update their representations in accordance to a dynamic world. The studies specifically explored how reasoning, and therefore judgment is linked to these representations. The thesis discusses the theoretical implications of these and other findings for the causal model framework as well as for cognitive science more generally. Related practical implications include the importance of understanding naïve causal models for applied fields such as legal decision-making and environmental psychology.

To my parents

Acknowledgements

Foremost, I would like to express my deepest gratitude to my supervisor Dave Lagnado. I consider myself exceptionally lucky to have had the privilege of being supervised by somebody who has been the greatest inspiration both as an academic and as a person. I thank him for his infinite guidance, help and understanding. I could not have imagined having a better supervisor and mentor for my Ph.D. study.

I would also like to thank all the members of my lab for always being there to give me valuable feedback. In particular, I would like to thank Adam Harris, Tobias Gerstenberg, Christos Bechlivanidis , Chris Olivola, Anne Hsu and Cristina Miclea for all their time and help through my research. Special thanks to Professor David Green, for his time consulting on some of the integral ideas in this thesis.

Finally, I would like to thank all of my amazing family. Thanks to my parents, Mamma and Pappi, for always being there for me in every possible way. Thanks to my sisters, Vali and Irene, for all their understanding throughout these last four years. Thanks to my grandparents, Dodo and Nonna. Thanks to my little friend Zen.

Table of Contents

Chapter 1: Introduction

The introduction starts by presenting the question that has driven the research reported in this thesis. The what, the why and the how are discussed to provide a brief idea of the rationale behind the question as well as the chosen approach. This is followed by an overview of the general background that forms the basis of the studies discussed in the ensuing three chapters. Finally, a brief outline describes the research question of each study.

1.1 The question

What?

The central premise of cognitive science is that thinking can best be understood in terms of representational structures in the mind and computational procedures that operate on those structures. This hypothesis can be framed into one question: "How do we represent our world and how do we use these representations to reason about it?" In a sense, this question encompasses two enquiries within one: the question of representation and the question of reasoning. These stand by themselves, but they are intrinsically interrelated in every way. Arguably, there cannot be one without the other. This thesis focuses exactly on investigating the nature of this interrelation and how it may shape judgment and decision-making in different aspects of everyday life.

Why*?*

The dominant tenet behind this question is that the quest of answering it yields important theoretical as well as practical implications. From a theoretical standpoint,

the question of representation and reasoning lies at the very core of most theories of how the mind works. This means that attempting to further comprehend how people understand the world and make decisions based on this understanding, will inform current theoretical accounts. From a practical perspective, there is much to be gained from adopting an applied approach. Understanding the driving forces that underlie people's decisions and actions can potentially help unlock solutions to many of today's complex problems (White, 2000).

How?

The question of how people represent the world and reason with these representations is complicated by the fact that the world is always changing. The very course of Nature is about change on every level. Everything with a physical realization is transient: the way in which seeds sprout, species evolve and oceans warm are just a few simple examples. Therefore reasoning and representation cannot possibly be based on these unstable states. Rather, it must be based on the idea that events do indeed change but the forces of change do not – they remain invariant across spatial and temporal contexts.

These forces of change are simply mechanisms of cause and effect. The temperature of the ocean is always changing - it is unstable. However, the causal relations that govern the mechanisms by which it warms up are invariant: more carbon dioxide in the atmosphere will always cause the ocean to warm up. This concept is embodied in the principle of causal invariance (Sloman 2009; Woodward, 2000). Accordingly, the causal relations that govern mechanisms of change form a reliable basis for general knowledge. After all, science is concerned precisely with discovering and representing causal structure - be it how force changes acceleration or

how overfishing changes ecosystems. The point is that the casual principles that govern mechanisms are useful because they apply across time and across a large number of objects – in other words, they allow generalization of empirical knowledge. Hence, it follows then that the logic of causality is people's guide to prediction, explanation and action. This logic is the tenet of the causal model framework (Pearl, 2000; Spirites, Glymour and Scheines, 1993) and will be used to approach the question of reasoning and representation.

1.2 Representation and reasoning

There have been substantial advances in normative models of both representation and reasoning over the past decade, and a variety of network models have been developed. Network models are simply graphical representations of causal relations (links) between events or causes (nodes). These include Wigmore charts (Wigmore, 1913), cognitive maps (Axelrod, 1976), and Bayesian networks (Pearl, 1998). Given that the central question concerns reasoning as well as representation, the only relevant network model is Bayesian networks. This is because a significant feature of Bayesian networks is that once the representation is constructed, it can be used for inference. This sets it apart from most other forms of networks (e.g. Wigmore charts and Cognitive maps), which serve mainly as descriptive tools. Indeed representation is intertwined with inference in a Bayesian network (Lagnado, 2011).

Bayesian networks are directed acyclic graphs in which the nodes represent the variables relevant to the situation (e.g., the presence of an alibi, the number of fish in the sea, the occurrence of an event) and the links represent causal relations among these variables. The strength of a causal relation is defined by conditional probabilities that are related to each collection of parents–child nodes in the network.

Figure 1 shows an example of a Bayesian network based on a classic example adapted from Pearl (2000).

The network in the figure describes the causal relationships among the following variables; season of the year (season), whether rain falls (rain) during the season, whether the sprinkler is on (sprinkler) during that season, whether the pavement would get wet (wet), and whether the pavement would be slippery (slippery). In this example, the lack of a direct link between 'season' and 'slippery' captures the idea that changes in season affect the slipperiness of the pavement by other intermediate factors (e.g., how wet the pavement is).

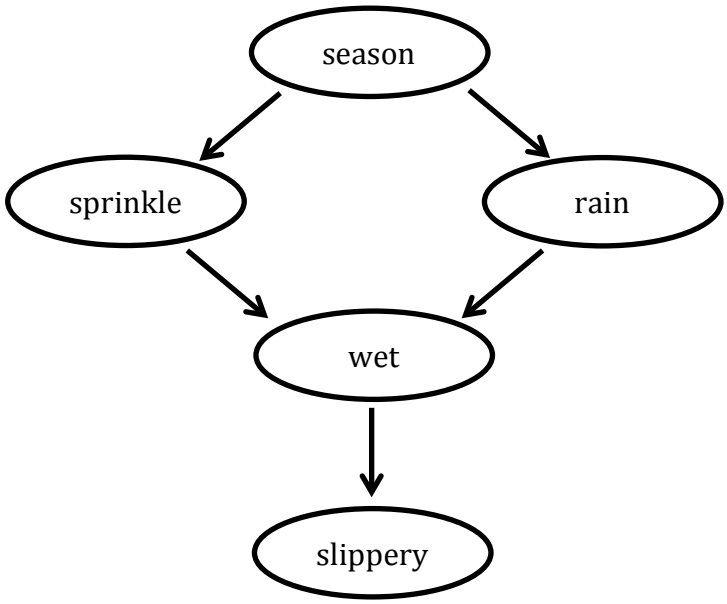


Figure 1

A Bayesian network representing causal influences among five variables.

Perhaps the most important thing to note in this example is that a Bayesian network models the environment as opposed to modeling a reasoning process. This is not the case in many other knowledge representation schemes like logic, rule-based systems, and neural networks. The fact that a Bayesian network simulates the causal

mechanisms in the environment means that it allows people to represent the world as it is. This implies that they can then answer a range of questions based on this representation. Such questions might include abductive questions, such as "What is the most plausible explanation for the pavement being wet?" and control questions, such as "What will happen if we switch off the sprinkler?" It is clear that the answers to these questions are based on the causal knowledge that can effectively be represented and processed in Bayesian networks.

Bayesian networks have well-established foundations in probability theory, and are currently applied in many practical contexts (e.g. medical diagnosis; Heckerman, 1991). Evidently, this formal approach is just prescriptive in the sense that it does not necessarily reflect the psychological reality of representation and reasoning. People do not always represent probability distributions accurately and do not always make sound probability judgments as set by causal model theory (Gilovich, Griffin and Kahneman, 2002; Kahneman, Slovic and Tversky, 1982). The question of how people really represent and reason necessitates a descriptive approach instead – an approach based on the causal model framework as a psychological theory (Sloman, 2005). This is not to say that a descriptive approach completely discards the prescriptive causal model theory. On the contrary, it can very much be based on some of the same core ideas. What follows is a brief description of some (by no means all of them) of these core ideas. These concepts shape the psychological theory of representation and reasoning that will guide the investigation of the central question put forward by the current thesis.

Causal networks

The first of these ideas concerns the structure of people's representations - that

is that people's representations are in the form of a network. This is the intuitive notion that individual cause-effect relations are not isolated from each other but tend to be understood in an organized representation of chains and networks of causal relations (White, 2008). This network arrangement lets people deal with complex multivariate causal reasoning: reasoning based on numerous interrelated pieces of information. This is because the whole is more than the sum of its parts. In other words, the structure of a causal belief system (a set of interrelated causal relations) is more informative than the isolated individual beliefs (Waldman & Hagmayer, 2005).

For example, if a tree was to be reduced to its individual parts (leaves, branches, trunk, bark, roots, fruit, and so on) it would not be possible to represent the whole tree's significance, such as the role the tree plays as habitat for birds, insects, parasitic vines, and other organisms. Similarly, chemically analyzing of the tree's chloroplasts, diagramming its branch structure, and evaluating its fruit's nutritional content, would not lead to understanding the tree as habitat, as part of the forest landscape, or as a reservoir for carbon storage.

Hence, the idea is that a cause, or variable within the network, can only be evaluated meaningfully with respect to its relation to other items represented within the network. This is not to say that people do not isolate small fragments of the network they represent - indeed this is what makes the whole network representation tractable for the human mind (Lagnado, 2011).

Qualitative relations

The second idea that provides the key to making causal networks tractable for everyday representation and reasoning is that people represent the qualitative structure of causal systems without actively representing all the quantitative details

(Wellman & Henrion, 1993). For example, a link from A to B tells us that certain values of A will change the probability of certain values of B, without needing to specify exactly how much.

Indeed, even the network structure underpinning a Bayesian Network is purely qualitative in the sense that it represents the presence or absence of a dependency between a set of variables. Even though the standard Bayesian Network framework requires a precise set of conditional probabilities, many of the important characteristics of the network are retained without a full and exact set of probabilities (Biedermann and Taroni, 2006; Wellman and Henrion, 1993). This means that even if people are unable to perform exact Bayesian computations over this network, they can still draw approximate inferences (perhaps using heuristic methods). There is growing empirical evidence that people reason in accordance with the qualitative prescripts of causal Bayesian Networks (Krynski and Tenenbaum, 2007; Sloman and Lagnado, 2005).

This idea is particularly significant in domains where no precise figures are available or where much of the information does not admit of quantification (there might be large numbers of interacting variables so that exact inference is intractable). For example, as Lagnado points out (2011) it might not be possible to quantify the exact probative force of a witness testimony that places the defendant at the crime scene; but most people would agree that it raises the probability of guilt, however slightly. Moreover, people will often be able to make comparative probability judgments; for instance, judging that a certain piece of forensic evidence raises the probability of guilt more that the testimony of a partial witness.

The idea that people utilize qualitative networks is not new. There are several psychology studies that speak in favour of qualitative approaches. First,

psychophysical studies show that for a range of sensory phenomena people are poor at making absolute judgements, and instead make ordinal comparisons (e.g. Stewart, Brown and Chater, 2005). Second, analyses of a wide range of predictive tasks (e.g. clinical and medical diagnosis) suggest that statistical models that use unit weights often outperform more complex models (Dawes, 1979). The key requirement for these simpler models is that the sign of each variable in the model is correct, while the exact weights placed on these variables is not significant. The proposed qualitative networks however, go beyond simple linear models, but share the intuition that precision in the weights is not a necessary condition for successful inference.

Dynamic models

The third idea is based on the appreciation that people' representations must adapt to a changing environment (Osman, 2010). These changes in the environment might involve new information, new hypotheses and new goals (Lagnado, 2011). Most events in everyday life are not detected based on a particular point in time, but they can be described through multiple states of observation that yield a judgment of one complete final event. This implies that many learning experiences involve repeatedly learning about one variable over time (Rottman & Keil, 2012). For example, one might develop beliefs about the causal relationships between fatigue, insomnia, and stress, by observing a person who experiences these conditions wax and wane over time. This temporal dependency of the states between variables is what constitutes a dynamic model.

Indeed, Bayesian Networks can been adapted to model scenarios in which the states of variables are temporally dependent. In these Bayesian Dynamic Networks (DBNs), at each time period the state of a variable is determined both by the causal

relationships within that time period and also by prior states. An example of a DBN can be seen in Figure 2. Figure 2 shows the relationship between stress, insomnia and fatigue modeled as a causal chain whereby stress affects insomnia, which in turn affects fatigue.

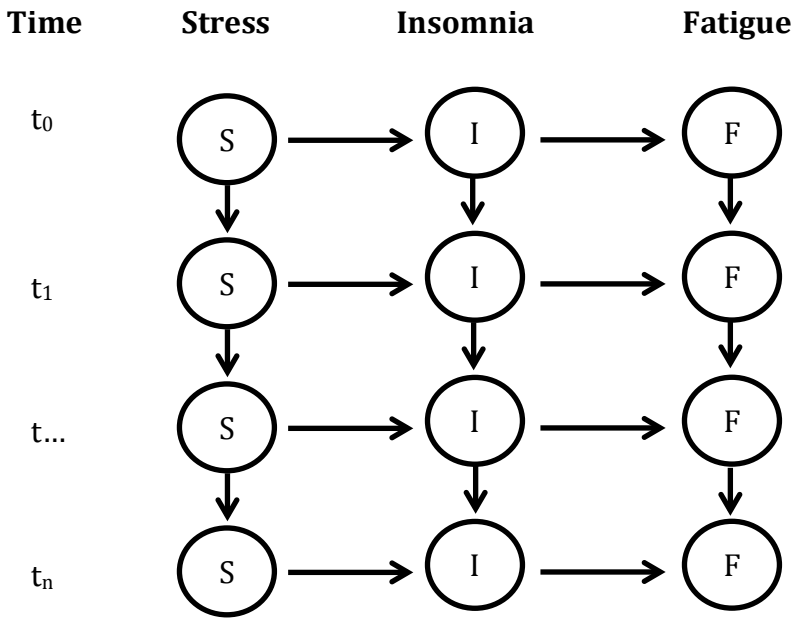


Figure 2.

An example of a dynamic causal model.

According to DBNs, the probabilities associated with each node are updated at each time period. This updating process involves intricate computations that can be executed by computational algorithms, sometimes using only approximate inference due to the computational complexity (e.g., Friedman, Murphy and Russell, 1998). As outlined above, people engage in a form of dynamic updating all the time. Naturally, the way in which they update, however, differs from the strictly Bayesian updating process in which a complete set of hypotheses are continually updated – that would be too demanding on a computational level. Rather, it has been proposed that people introduce and eliminate hypotheses in a more all-or-nothing manner (Lagnado, 2011)

– which makes inference considerably easier.

This concept fits well with the idea that people's representations are in the form of a network. This structure allows people to update the representation by recruiting only the relevant fragments of the network – the ones affected by the new information. Similarly, the qualitative nature of the representation means people can easily and readily update their representation without having to carry out numerous probability estimations and computations.

1.3 Outline

The current thesis investigates the answer to the question of reasoning and representation through three related yet distinct studies. All of them seek to explore the connection between representation and reasoning but in different forms.

Before outlining these three studies that form the basis of each of the following chapters, it is worth pointing out two key features they all have in common. The first of these is that they are all based on lay people. This is a crucial point because lay people have a way of representing and reasoning that differs from those of experts (Hoffman, 1996). This is true for most fields - the way a patient represents the causes of heart disease will undoubtedly differ from that of the doctor (Cuthbert, Dubolay, Teather et al., 1999). Similarly, even if lay people and experts were to have the same representations, the inferences drawn from them are bound to differ. Nevertheless, there is much to be learnt from lay people's representations and reasoning patterns – especially because they form the basis of general decision-making and therefore signify the power of the masses.

This leads to the second feature accompanying these studies: they are all carried out with practical applications in mind. This means that the investigations are

based on how people represent real world problems and how they might reason about them in an applied context. The applied approach starts off with Study 1, which is based on legal inference with a focus on juror decision-making. What makes the legal domain an appealing field of study is that many of its staple characteristics are also at the core of other significant applied domains. Evidence presented at trials tends to be contradictory, incomplete, biased and available in multiple formats that make them hard to compare and integrate. In addition the verdict needs to be reached under time constraints and limited cognitive resources. It is easy to see how these features are not exclusive to juror decision-making. These attributes are also the ones governing environmental problems, particularly from a consumer's point of view. The everyday consumer is constantly being fed contrasting evidence about climate change and anthropogenic effects of human behaviour. The information comes in many forms, from news reports, to word-of-mouth. Some of it is in the format of a statistic whilst other is simply a general opinion. Yet consumers have to make decisions that concern the environment everyday, every time they choose one product over another or chose to how to get to work. That is why the second study extends the exploration of representation and reasoning to the applied domain of environmental problems, dealing with one of the most complex contemporary issues of all: overfishing. The third study takes the investigation of the representation and reasoning to simpler domains including sleeping patterns and simple predator-prey relationships. This move to more familiar subjects was necessary to be able to study reasoning based on the more complex representations (causal loops) which drive the final study.

Study 1: Reasoning with causal evidence: understanding legal inferences.

The first study investigated how mock jurors represent criminal evidence and reason with causal network representations. Specifically, it explored: i) how jurors represent evidence (with a focus on how they update their representations in the face of new evidence); ii) the extent to which they arrange evidence in a causal network structure; and iii) jurors' ability to make causal inferences in line with their representations.

Study 2: Reasoning with causal networks: understanding environmental problems.

The second study examined how people represent the causes of a complex environmental problem (overfishing) and how their individual representations are directly linked to how they reason about the issue. In particular, it studied: i) whether people construct two-way causal relations (causal loops) within their representations; ii) the relation between the representations and counterfactual judgments; and iii) which features of their representations (e.g. causal strength of relations versus sheer number of causal relations) is most significant in predicting judgments.

Study 3: Reasoning with causal loops: understanding everything.

The third and final study inspected how people represent causal loops and reason in accordance to these cyclical representations. These are a form of simple qualitative dynamic networks. Specifically, the study employed different applied scenarios to survey the extent to which people can make proper causal inferences given the loops portrayed by these situations.

Chapter 2: Reasoning with causal evidence: understanding legal inferences

2.1 Introduction

The jury is the only decision-making body in the criminal justice system composed of laypersons and jury verdicts directly affect the implementation of justice in the USA, UK and elsewhere. This means that the lives of hundreds of thousands of individuals around the world every year depend on the fairness of jury decision-making. Researching the psychological processes underlying juror decision making, and the way these relate to formal methods of evidence evaluation, is therefore of fundamental importance to the criminal justice system. Unfortunately, in the quest to identify sources of bias in jurors' decisions, researchers have often overlooked the importance of approaching jurors' decision making from a cognitive perspective. As a result, little is known about the causal models underlying jurors' reasoning processes.

A jurors' task involves making decisions based on multiple pieces of probabilistic evidence. In addition these pieces of evidence tend to be contradictory, incomplete, biased and available in multiple formats which make them hard to compare and integrate. That being said, jurors nonetheless make meaningful decisions. Even though the exact mechanisms by which this happens still present an open question, their behaviour has been explained by descriptive models centred on the idea of sense making and constructing coherent stories from evidence (Pennington & Hastie, 1986). Most of the empirical studies conducted so far provide support for either the story model or the coherence model - two similar frameworks that have arisen from two very different traditions (Byrne, 1993). The story model comes from

the tradition of psychology of jury decisions which attempts to understand how it is that jurors arrive at a particular verdict. The coherence model on the other hand, derives from the tradition of philosophy of science, which attempts to understand how it is that scientists come to accept new paradigms. Interestingly, research beginning in these two seemingly disparate domains has converged in the sense that both approaches posit that evidentiary conclusions are not derived from mathematical computations of the independent values of raw evidence. Inferences, rather, are based on constructed representations of coherence, and it is these constructed representations that ultimately determine the verdicts (Bex, 2004).

The Story Model

The most widely cited and influential model of juror decision making is the story model. It has been proposed by Pennington and Hastie in 1981. It has received wide empirical support (e.g. Pennington & Hastie, 1981 1986, 1988, 1992, 1993) and it is still accepted as the standard in psychology and legal studies.

The story model proposes that jurors construct a narrative storyline out of the evidence presented during the trial. Pennington and Hastie (1992) suggest this happens over three stages: i) evaluating the evidence through story construction; ii) representing of the decision alternatives by learning the various verdict options available; and iii) reaching a decision by fitting the story to the most appropriate verdict category. During the story construction stage, jurors use three kinds of information to create a plausible story: i) the evidence presented throughout the trial; ii) personal knowledge about similar cases and iii) generic expectations about what makes a complete story.

Naturally, this story construction process can yield different interpretations of the evidence and hence may result in the construction of different stories. Pennington and Hastie (1992) propose that the criteria that jurors use to evaluate which story should prevail (or, in their words, elicit more acceptability and confidence) are the 'certainty principles'. These are composed by two main elements: *coverage* and, more importantly, *coherence*. Coverage simply refers to the extent to which the constructed story is able to provide an explanatory account of all the pieces of evidence. Coherence, on the other hand, is assigned a more prominent role in deciding which story is more acceptable. According to the story model, coherence is the product of three components: i) plausibility, ii) consistency, and iii) completeness. Therefore, Pennington and Hastie propose that a story with high coherence is a story that i) does not contain internal contradictions (high plausibility); ii) is consistent with events in the real world (high consistency) and iii) is complete (high completeness). Consequently, the story that will be evaluated to have greater coverage and coherence will be the story that will be deemed as more acceptable (and hence generating superior confidence).

The second stage of the story model is verdict representation. Information for verdict representation is given to jurors at the end of the trial. Jurors learn about the verdict options from the judge's instructions. However jurors may have pre-existing ideas about the meaning of verdict categories. Even though Pennington and Hastie argue that verdict representation is the second stage of the story model, they do not specify whether it happens in parallel with story construction (as jurors may refer to pre-existing schemas of verdicts when organising trial evidence) or once a story has been constructed and accepted.

In the final stage of the story model, jurors perform a matching task whereby they match the attributes of the story that they constructed in the first stage to the crime elements of one the verdict categories of the second stage. This task is moderated using the legal rules and prescriptions provided by the trial judge. In principle if the constructed story fits the requirements of the verdict category under consideration, the juror will choose that verdict category. If the threshold is not met, the juror will search for a more appropriate verdict category. Pennington and Hastie (1986) have tested their model by conducting many studies (Pennington & Hastie, 1981 1986, 1988, 1992, 1993) in which mock jurors are requested to carry out their deliberations out loud. As a result, they have presented substantial empirical evidence in support of the model. Nonetheless, one of the main limitations of the story model is that it is vaguely specified with respect to the underlying cognitive processes and mechanisms.

This problem is clear across different aspects of the model. First, as pointed out by Lagnado (2011), no precise account is given for how people update or construct their causal models, or how they draw inferences from them. Similarly, Harris and Hahn (2009) have argued that although Pennington and Hastie assign coherence a key role within this framework, they do not provide a formal way of formalising or measuring it. Furthermore the story model's path between story selection and determinations of guilt/culpability is unclear. This means that the story model does not provide clear insight into the juror's cognition and therefore is limited in its explanatory power. Following from this idea, coherence models have gained support as an alternative, but also complementary account of juror decision making (Simon & Holyoak, 2002; Simon, Snow & Read, 2004; Thagard, 2000).

Coherence models

The main idea behind coherence models is that the mind strives for coherent representations. The concept of coherence is at the centre of multiple frameworks within the domain of philosophical logic (e.g. Olsson, 1998). However, the formal approach to coherence that has gained more support across the realm of legal epistemology and more importantly, cognitive psychology, is one coming from the field of computational philosophy: Thagard's explanatory coherence model (1989). Two main features set Thagard's theory apart from other coherence theories. Firstly, by constructing a computational model, Thagard has provided a more detailed account of coherence-based reasoning than philosophers have traditionally done. The details, however, also allow us to see the problems of the coherence-based methodology it is premised on. Secondly, it has provided a general characterization of coherence as 'constraint satisfaction'.

Thagard (2006) listed seven principles that concisely state the theory of explanatory coherence. These are best illustrated by applying them to a criminal scenario. For example if a house has been burnt down, the police may consider the house owner and an arsonist as the potential suspects.

1. *Symmetry*. Symmetry refers to the idea that explanatory coherence is a symmetrical relation. That is, two propositions P and Q cohere with each other equally. This means that the hypothesis that the house owner burnt down the house coheres with the evidence that the house has been burnt down. This coherence relation is symmetrical as 'they hang together equally'.
2. *Explanation*. The principle of explanation is characterised by three propositions. The first one is that a hypothesis coheres with what it explains, which can either be evidence or another hypothesis. This refers to the fact that the hypothesis that

the house owner burnt down the house explains the evidence that the house is burnt down, so the hypothesis and the evidence cohere with each other. The second proposition of the principle of explanation is that hypotheses that together explain some other proposition cohere with each other. This idea allows for hypotheses to explain each other. For example, if there is the hypothesis that the house owner burnt down the house, this hypothesis can be explained by the hypothesis that he had a motive, for example that he was in financial trouble. There can even be multiple motives, for instance that the house owner was in financial trouble and depressed, and both of these hypotheses cohere with each other. Finally there is a third proposition based on the idea of simplicity. That is the more hypotheses it takes to explain something, the lower the degree of coherence. Simplicity is a matter of explaining a lot with few assumptions.

3. *Analogy*. Analogy is the principle that similar hypotheses that explain similar pieces of evidence cohere with each other. For example, if the house owner had a history of financial trouble and depression resulting in destroying goods to claim insurance, then these cases provide analogies that the house owner did it more plausible in the current case.
4. *Data priority*. The fourth principle refers to the idea that propositions that describe the results of observations have a degree of acceptability on their own. For example there can be an observation that the house owner had petrol traces on his clothes. This observational evidence would get a degree of coherence on its own, providing a degree of priority to such observations. It is important to keep in mind that this principle does not require the observations to be indubitable but leaves open the possibility that explanations could be found to be erroneous despite their initial degree of coherence.

5. *Contradiction*. Contradictory propositions are incoherent with each other. This refers to the straightforward case in which two hypotheses are logically contradictory: for example the hypothesis that the house owner did it contradicts the hypothesis that the arsonist did it, then these two hypotheses are incoherent.
6. *Competition.* Competition refers to the idea that if P and Q both explain a proposition, and if P and Q are not explanatorily connected, then P and Q are incoherent with each other. The hypothesis that the house owner did it competes with the hypothesis that the arsonist did it. Since these two hypotheses independently explain evidence they are treated as competitors that are incoherent with each other. However there could be circumstances whereby it could be logically possible for the house owner and the arsonist to have burnt down the house together. At that point if there was reason to believe that the house owner and the arsonist acted together in a conspiracy, then the two hypotheses would be explanatorily connected and would be treated as coherent with each other.
7. *Acceptance*. The last proposition proposes that the acceptability of a hypothesis in a system of hypotheses depends on its coherence with them. In other words, hypotheses should be accepted and rejected on the basis of their overall coherence with each other. Because these hypotheses can be coherent and incoherent in many ways, acceptability makes inference a highly complex and nonlinear process. For this reason explanation evaluation is executed through simple artificial neural networks.

These seven principles do not fully specify how to determine coherence-based acceptance, but algorithms are available that can compute acceptance and rejection of propositions on the basis of coherence relations. The most psychologically natural algorithms use artificial neural networks that represent propositions by artificial

neurons or units and represent coherence and incoherence relations by excitatory and inhibitory links between the units that represent the propositions. Acceptance or rejection of a proposition is represented by the degree of activation of the unit. The program ECHO spreads activation among all units in a network until some units are activated and others are inactivated, in a way that maximizes the coherence of all the propositions represented by the units (see Thagard, 1992, 2000 for the technical details). Several different algorithms for computing coherence are analyzed in Thagard and Verbeurgt (1998).

In the crime example, the hypothesis that the house owner burnt down the house can be represented by a unit called HOUSE OWNER DID IT and the evidence that the house is burnt down by a unit called HOUSE BURNT. Then, whenever principles of explanation and analogy establish relations of coherence between two propositions, the units that represent the propositions get excitatory links between them. Thus HOUSE OWNER DID IT and HOUSE BURNT have an excitatory link between them that is symmetrical (in accord with principle of symmetry). The principle of data priority is implemented by making an excitatory link between the special unit EVIDENCE and any unit such as HOUSE BURNT that represents a proposition based on observation. The principles of contradiction and competition, which establish incoherence between competing hypotheses, are implemented by means of inhibitory links between units: When two hypotheses are incoherent—e.g., the house owner did it versus the arsonist did it—then the units that represent the hypotheses—HOUSE OWNER DID IT and ARSONIST DID IT—will get an inhibitory link between them.

Figure 1 depicts the simple network that evaluates competing explanations in the house fire example. It includes a unit called OWNER IN FINANCIAL

TROUBLE that represents the hypothesis that the house owner was in financial trouble, and a unit called OWNER HAS PETROL TRACES that represents the evidence that the house owner had petrol traces on his clothes. The excitatory links between units representing coherent propositions and the inhibitory links between units representing incoherent propositions. In simulations, the links have different weights that can represent the degree of coherence or incoherence between propositions.

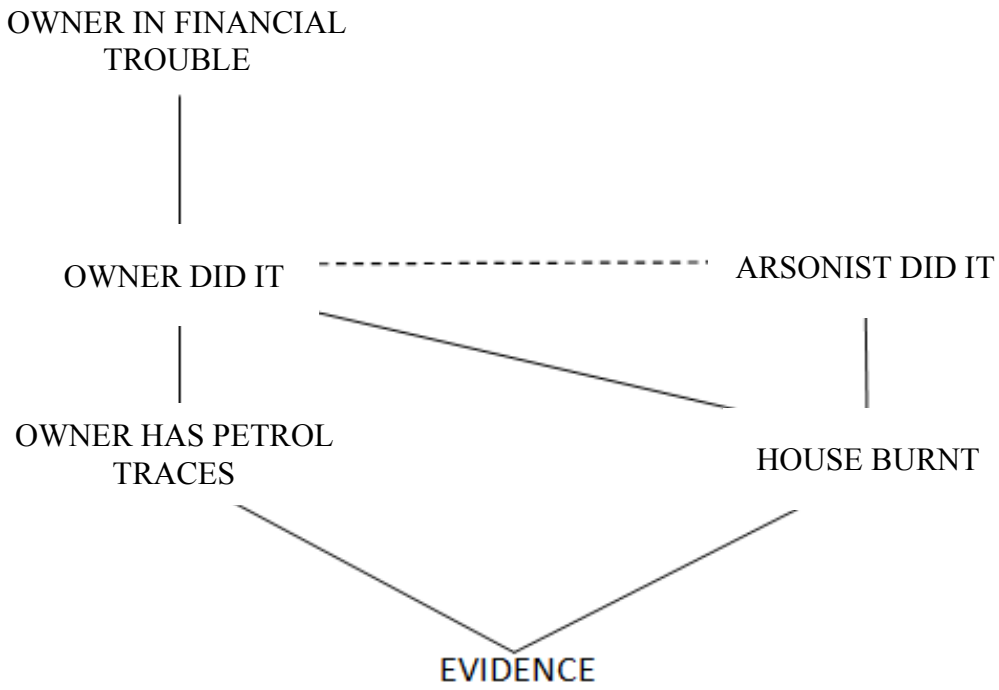


Figure 1. Neural network modelling competing explanations for the burning down of a house. The straight lines indicate coherence relations (positive constraints) established because a hypothesis explains a piece of evidence. The dotted lines indicate incoherence relations (negative constraints).

Thagard's theory (through ECHO) has been used to model some prominent jury verdicts (e.g. Thagard, 1989) and does much by way of solving some of the problems which beset coherence theories of justification, including coherence theories of legal justification (see Amaya, 2007). However, it still has a fundamental limitation. Lagnado (2011) has argued that coherence models are unable to represent basic forms of inference such as 'explaining away' (Pearl, 1988).

Explaining Away

Explaining Away is a common and intuitively compelling pattern of inference that refers to the idea that because one cause explains the observed effect it therefore reduces the need to invoke other causes. Wellman and Henrion (1993) illustrate this with the following example. A friend sneezes and this raises the probability of him having a cold, and the probability of him having an allergic reaction. Once it is found out that the friend is allergic to cats, and a cat is observed to be present, this lends confirmation to the hypothesis he is having an allergic reaction. This explains away the sneezing and, therefore, reduces the probability of the cold. In other words the two hypotheses were independent when the status of the evidence was unknown, but become conditionally dependent given its status.

This pattern of inference is naturally captured using a Bayesian Network representation. Bayesian networks have well-established foundations in probability theory, and are currently applied in many practical contexts. Bayesian networks consist of two parts: a graph structure and a set of conditional probability tables. The graph structure is made up of a set of nodes corresponding to the variables of interest, and a set of directed links between these variables corresponding to causal relations. The variables tend to be causes and effects but they could even be hypotheses about

pieces of evidence (such as in legal contexts). This yields a directed graph that represents the probabilistic relations between variables, in particular the conditional and unconditional dependencies. In addition to the graph, a Bayesian network also requires a conditional probability distribution table for each variable. This dictates the probability of the variable in question conditional on the possible values of its parents (the nodes with direct links into that variable). This arrangement of nodes and links, plus the conditional probability tables for each node, dictate what inferences are licensed (via the laws of probability).

In a simplified model, there are three binary variables, *Cold* (C) which represents whether or not someone has a cold, *Allergy* (A) which represents whether or not someone has an allergy, and *Sneeze* (S) which represents whether or not someone sneezes. Both C and A are potential causes of S. The graph structure is depicted in Figure 2. This encodes the assumption that C and A are marginally independent, i.e., P(C) = P(C|A). The observation of sneezing raises the probability of both C and A: P(C|S) > P(C); P(A|S)>P(A). However, on observing that A is true, the probability of C returns to its prior level: P(C|A&S)=P(C). This is the basic phenomena of explaining away.

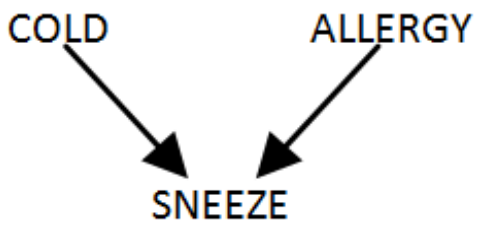


Figure 2. Graph structure for simple example of explaining away.

To understand how coherence models cannot cope with such a causal network it is sufficient to look at the way they would represent a simple criminal scenario.

Returning to the 'house fire' example, there could be two potential explanations for the petrol traces evidence: i) the petrol traces are from the petrol used to start the fire, or ii) the petrol traces are the result of the house owner spilling petrol while filling up his car. According to the first explanation the probability that the house owner is guilty is raised whilst according to the second explanation the probability of guilt is lowered. General coherence models assume that if two propositions both explain a piece of evidence, but they are not explanatorily connected, then the two propositions are incoherent with each other. This is expressed explicitly in Thagard's 'competition' principle: if P and Q both explain a proposition, and if P and Q are not explanatorily connected, then P and Q are incoherent with each other.

Returning to the house fire scenario example, the hypothesis that the owner spilled petrol while filling up his car, and the hypothesis that the owner has petrol traces on clothes because he started a fire, independently explain evidence and would therefore be treated as competitors that are incoherent with each other. However, this is an inappropriate representation of their true relation in the world. Whether or not the suspect spilled petrol while filling up his car is unrelated (independent) of whether or not he is guilty of starting the fire. The two explanations only become dependent given the evidence of petrol traces on the clothes that they both try to explain. This is clear by looking at Figure 3. It depicts the neural network that evaluates competing explanations for the petrol traces, as modelled by coherence accounts. Solid lines are excitatory links between units, and the dotted line is an inhibitory link representing incoherence between competing hypotheses about the origin of the petrol traces—starting a fire or spilling it while filling up car.

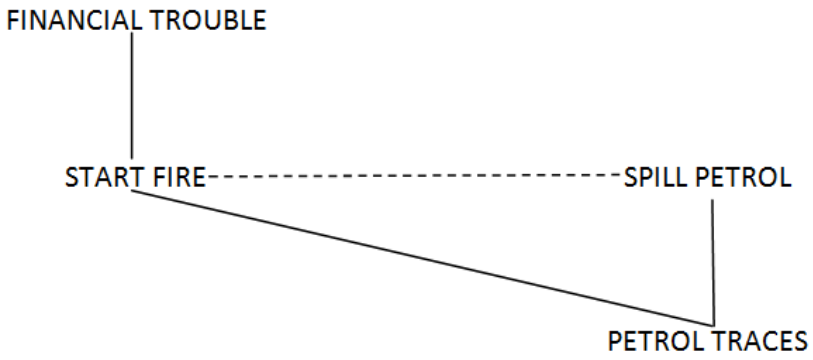


Figure 3. Connectionist network modelling competing explanations for petrol traces.

2.2 Experiment 1

The aim of this study was to show that mock jurors reason through crime scenarios by constructing causal networks that allow them to perform explaining away inferences that cannot be accounted for by coherence models. This was done by presenting participants with two fictional criminal scenarios based on a simple causal network that involved an explaining away causal inference.

Both scenarios began with background details about the crime and the chief suspect. Participants were then asked to make two baseline judgments. The first one was an estimate of the likelihood that the suspect was guilty (guilt judgment). The second one was an estimate of the likelihood of another event that might have caused the suspect to commit the crime (causal judgment). Once participants made these judgments they were presented with a new piece of evidence that incriminated the suspect (affirmative evidence) and were asked to make the same two judgments again. Lastly, they were presented with a new piece of evidence that explained away the previous piece of affirmative evidence (rebuttal evidence). Finally participants were asked to make the same judgments again. Participants thus gave six sequential

probability judgments in each problem (two baseline, two after affirmative evidence and two after rebuttal evidence).

This causal network is best explained using one of the crime scenarios as an example. The first scenario provided participants with the following background information: 'A house burnt down. The police investigation reveals that the fire was caused by ignited petrol and was therefore intentional. Statistics show that in about half of cases of intentional burning down of houses, it is the owner who is responsible for starting the fire to get compensation money from insurance.' Participants were then asked to make two baseline judgments. The first one was an estimate of the likelihood that the house owner was guilty (guilt judgment). The second one was an estimate of the likelihood that the owner was in financial trouble (causal judgment). After participants made their judgments they were provided with new affirmative evidence: 'Further police investigation revealed that the owner had petrol traces on his clothes.' Participants were then asked to make the same judgments as before. Next, participants were presented with a piece of rebuttal evidence which 'explained away' the piece of affirmative evidence: 'Further police investigation revealed CCTV footage of the owner filling up his car with petrol and spilling petrol on his clothes.' Finally participants were asked to make the two judgments one last time.

In contrast to the network representation suggested by the coherence approach (Figure 3), participants were hypothesised to have constructed a causal model that allowed for 'explain away' inferences. This is displayed in Figure 4.

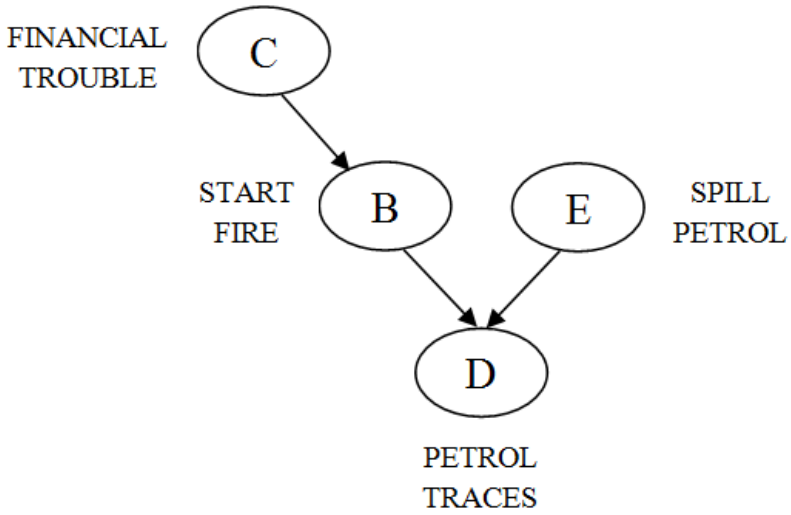


Figure 4. Causal network for house fire scenario.

The experimental hypotheses were:

i) guilt judgments after the affirmative evidence will be greater than the baseline judgments, showing that the affirmative evidence did in fact have the intended incriminating impact;

ii) causal judgments will also be greater after the affirmative evidence as participants create a causal link between the evidence and the causal judgment;

iii) guilt judgments after the rebuttal evidence will be lower than the judgments given after the affirmative evidence judgments, showing that participants make explaining away inferences;

iv) causal judgments will also be lower than the judgments given after the affirmative evidence because participants extend explaining away inferences to causal judgments.

The causal network in question and the relative hypotheses can also be expressed probabilistically (note that letter 'A' stands for the background evidence):

i) $P(B|A) > P(B| A\&D)$;

ii) $P(C|A) > P(C| A\&D)$;

iii) $P(B| A \& D) > P(B|A \& D \& E)$;

iv) $P(C| A \& D) > P(C|A \& D \& E)$.

In order to assess whether participants had constructed the hypothesised causal model, they were presented with a series of questions that aimed to assess their causal schemas. In accordance it was hypothesised that participants would represent:

i) the guilt variable to be positively linked to the causal variable;

ii) the guilt variable to be positively linked to the affirmative evidence;

iii) the rebuttal evidence to be positively linked to the affirmative evidence

iv) No link between the causal variable and the rebuttal evidence.

Importantly, the fourth hypothesis (stating the absence of a causal link between the casual variable and the rebuttal variable), directly assessed whether participants constructed an inhibitory link between the two alternative explanations for the evidence. Results supporting the experimental hypothesis would speak strongly against a coherence model based representation of the evidential reasoning.

2.3 Method

Participants and Apparatus

65 first year undergraduate students from UCL (University College London) participated in the study in return for course credit. 52 participants were female and the mean age was 18.9 (1.47). The experiment was conducted online on individual computers and programmed in Adobe Dreamweaver.

Design

The experiment followed a within-subject design where each participant was presented with both scenarios. The order of the two scenarios was counterbalanced.

Materials and Procedure

The materials consisted of two scenarios. : the 'House fire scenario' and the 'Injured child scenario'. Each scenario was accompanied by a set of judgment questions and ended with four causal questions

Scenarios and judgment questions

These are reported in Table 1 along with the questions, in the same order they were presented during the study. Both scenarios began with background details about the crime and the chief suspect. This simple description was followed by two questions: the baseline guilt judgment and the baseline causal judgment. Participants indicated their judgments on a slider scale that was labeled as from 'extremely unlikely' to 'extremely likely'. The label in the center read 'as likely as not'. This is shown in Figure 5. The scale did not have any numbers, but the responses were coded from 0 to 100 (with 0 = extremely unlikely, 50 = as likely as not, and 100 = extremely likely).

Figure 5. Response scale for the guilt and causal judgments.

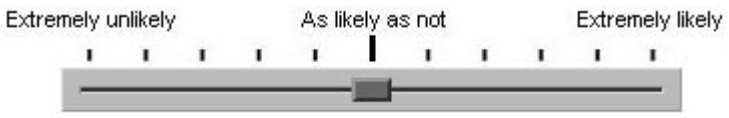


Table 1.

Table displaying the material shown to participants in the order in which it was presented. Experiment 1.

Scenario	House fire	Child
Background details	A house burnt down. The police investigation reveals that the fire was caused by ignited petrol and was therefore intentional. Statistics show that in about half of cases of intentional burning down of houses, it is the owner who is responsible for starting the fire to get compensation money from insurance as they are in financial trouble.	A child who has been brought into hospital has serious head trauma related injuries. Medical and police records show that about half of children with those specific injuries have been violently shaken by their parents.
Baseline guilt judgement	Please indicate how likely it is that the owner was the one who burnt down the house.	Please indicate how likely it is that the child has been violently shaken.
Baseline causal judgment	Please indicate how likely it is that the owner was in financial trouble.	Please indicate how likely it is that the parents have mental health problems.
Affirmative evidence	Further police investigation revealed that the owner had petrol traces on his clothes.	Further police investigation revealed that the injuries presented by the child included retinal haemorrhages – a characteristic symptom of violent shaking.
Affirmative guilt judgement	In light of this new evidence please indicate how likely it is that the owner was the one who burnt down the house.	In light of this new evidence please indicate how likely it is that the child has been violently shaken.
Affirmative causal judgment	In light of this new evidence please indicate how likely it is that the owner was in financial trouble.	In light of this new evidence please indicate how likely it is that the parents have mental health problems.
Rebuttal evidence	Further police investigation revealed CCTV footage of the owner filling up his car with petrol and spilling petrol on his clothes.	Further police investigation revealed that the child was born with a severe vitamin C deficiency which causes retinal haemorrhages.
Rebuttal guilt judgement	In light of this new evidence please indicate how likely it is that the owner was the one who burnt	In light of this new evidence please indicate how likely it is that the child has been violently shaken.
Rebuttal causal judgment	In light of this new evidence please indicate how likely it is that the owner was in financial trouble.	In light of this new evidence please indicate how likely it is that the parents have mental health problems.

Participants were then presented with a piece of affirmative evidence that incriminated the suspect. Following this new information they were asked to repeat the two judgments (affirmative guilt judgment and affirmative causal judgment). Next they were presented with a piece of rebuttal evidence that discredited the affirmative evidence – it explained it away. Finally participants were asked to make the last two judgments: rebuttal guilt judgment and rebuttal causal judgment.

Causal questions

Four questions were constructed for each scenario to assess the causal model participants had constructed. Each question consisted in a forced choice question (these are reported in Table 2) and a confidence judgment. The confidence judgment consisted in a likelihood rating.

Table 2.

Table showing the causal questions for each scenario. Experiment 1.

Scenario	House fire	Child
Causal question 1	Do you think knowing a suspect might be in financial trouble is relevant to whether the suspect burnt down his house?	Do you think knowing that suspects might have mental health problems is relevant to whether the suspects are responsible for violently shaking their child?
Causal question 2	Do you think knowing a suspect has petrol traces on his clothes is relevant to whether the suspect burnt down his house?	Do you think knowing a child has retinal haemorrhage is relevant to whether the child has been violently shaken?
Causal question 3	Do you think knowing a suspect has spilt petrol on his clothes while filling up his car is relevant to knowing the suspect has petrol traces on his clothes?	Do you think knowing a child has vitamin C deficiency is relevant to knowing the child has retinal haemorrhage?
Causal question 4	Do you think knowing a suspect might be in financial trouble is relevant to knowing the suspect has spilt petrol on his clothes while filling up his car?	Do you think knowing suspects might have mental health problems is relevant to knowing their child has vitamin C deficiency?

The reason why the 'Equally likely option' was not given as a possible answer in the forced choice question, was to tackle the worry that participants might select it to remain neutral (avoid giving a wrong answer) rather than to indicate actual feeling that the connections are indeed equally likely. For this reason the forced choice question was followed by a confidence judgment. It was presumed that participants who constructed a causal link would give a higher likelihood rating whereas participants who did not construct a causal link (i.e. thought the two answers were equally likely) would select one of the answers and then give the lowest likelihood rating. For example, to assess the presence of a positive causal link between the guilt variable and the casual variable in the 'Injured child scenario', participants were asked:

Which parents are more likely to violently shake their child?

- Parents with mental health problems
- Parents without mental health problems

How much more?

Figure 6. Response scale for the causal questions

2.4 Results

Scenario judgments

The mean and standard deviation of the three sets of guilt and causal judgments are displayed in Table 3.

Table 3.

Mean and standard deviation of guilt judgments and causal judgments. Experiment 1.

	House fire scenario		Injured child scenario	
Judgment	Guilt	Causal	Guilt	Causal
Baseline	51.4 (9.8)	52.9 (11.6)	51 (10.3)	47.3 (16)
Affirmative	69.8 (12.4)	63.5 (13.6)	73.2 (12.3)	59.7 (19.9)
Rebuttal	46.9 (14.3)	48.1 (14.6)	41.7 (13.4)	42 (13.8)

T-tests

Paired samples t-tests were conducted to evaluate the significance of the results. The way the judgments change after the affirmative and rebuttal evidence is clear by observing Figure 6 (House fire scenario) and Figure 7 (Injured child scenario)

House fire scenario. The first hypothesis was supported as there was a significant increase between guilt judgments given at baseline and those given after the affirmative evidence: $t(64)= -8.89$, $p<0.001$. The same was true for causal judgments (hypothesis 2): $t(64)= -4.48$ $p<0.001$. Also supported was the third hypothesis, which predicted that participants' guilt judgments given after the rebuttal evidence would be significantly lower than the ones given after the affirmative evidence: $t(64)=9.151$, $p<0.001$. The same was true for causal judgments (hypothesis 4) $t(64)=6.144$, $p<0.001$.

Interestingly, inspection of the means and graph showed participants' guilt and causal judgments given after the rebuttal evidence were both lower than the initial baseline judgments. A paired sample t-test was conducted to evaluate whether there was a significant difference between the rebuttal guilt judgment and the baseline guilt judgment. The result was indeed significant, $t(64)=2.218$, $p=0.03$. The same was true for the difference between the rebuttal causal judgment and the baseline causal judgment, $t(64)=2.18$, $p=0.033$.

Injured child scenario. The first hypothesis was supported as there was a significant increase between guilt judgments given at baseline and those given after the affirmative evidence: $t(64)= -10.8$, $p<0.001$. The same was true for causal judgments (hypothesis 2): $t(64)= -3.9$ $p<0.001$. Also supported was the third hypothesis, which predicted that participants' guilt judgments given after the rebuttal evidence would be significantly lower than the ones given after the affirmative evidence: $t(64)=14.9$, $p<0.001$. The same was true for causal judgments (hypothesis 4) $t(64)=5.54$, $p<0.001$.

Interestingly, inspection of the means and graph showed participants' guilt and causal judgments given after the rebuttal evidence were both lower than the initial baseline judgments. A paired sample t-test was conducted to evaluate whether there was a significant difference between the rebuttal guilt judgment and the baseline guilt judgment. The result was indeed significant, $t(64)=-4.286$ $p<0.001$. On the other hand, the difference between the rebuttal causal judgment and the baseline causal judgment was not signficiant, $t(64)=-1.86$, $p=0.067$.

Figure 6. Graph showing the mean guilt and causal judgments (with error bars). Experiment 1, 'House Fire' scenario.

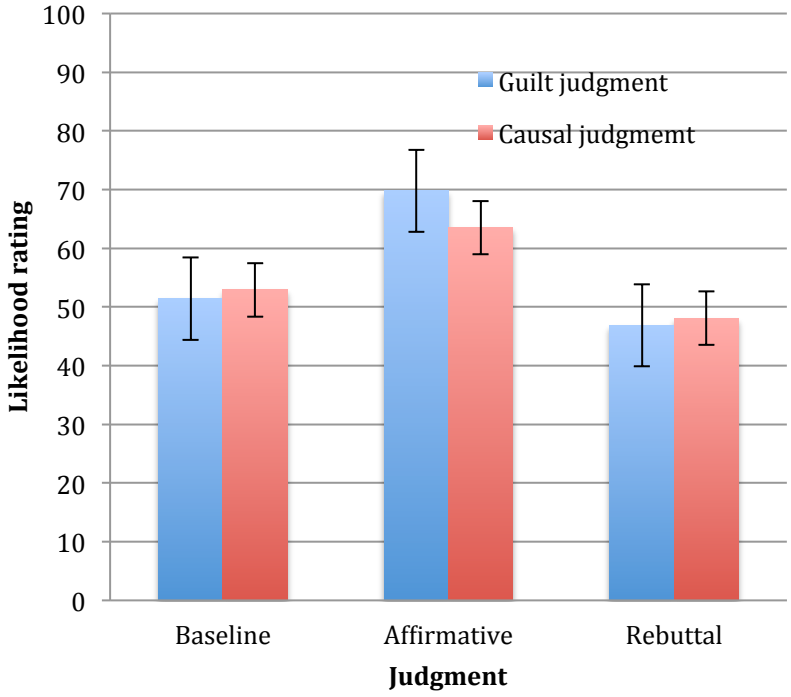
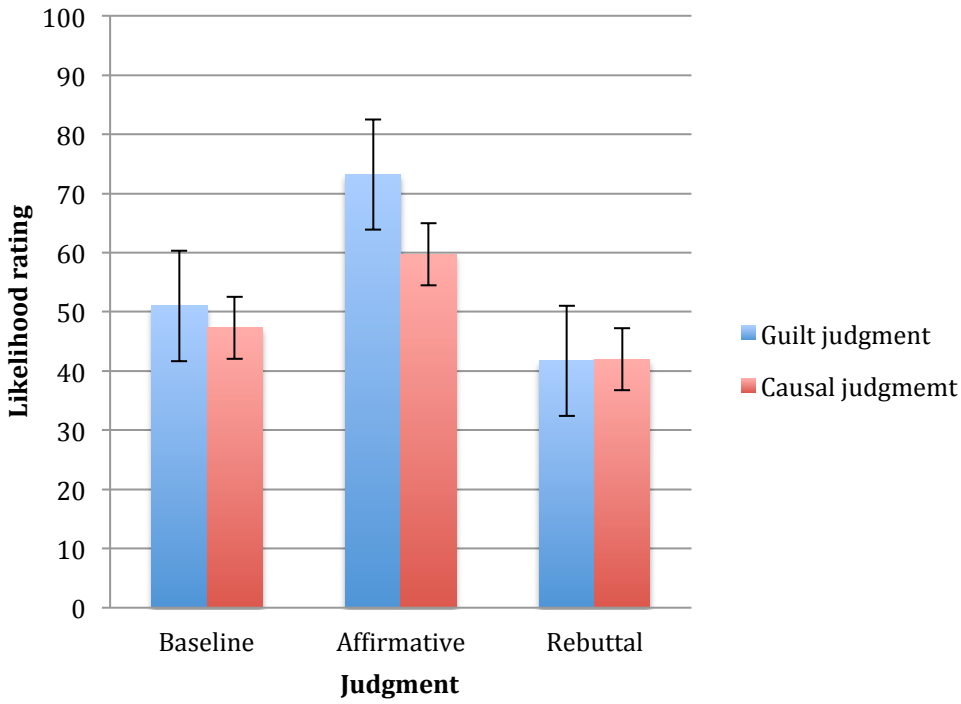


Figure 7. Graph showing the mean guilt and causal judgments (with error bars). Experiment 1, 'Injured child' scenario.



ANOVAs

House Fire scenario. A one-way repeated measures ANOVA was conducted to compare the guilt judgments given at the three different stages (after background evidence, after affirmative evidence and after rebuttal evidence). There was a significant effect for guilt judgment, Wilk's Lambda = 0.388, F (2, 63) = 49.6, $p < 0.001$, multivariate partial eta squared = 0.61. Post hoc tests using the Bonferroni correction revealed that, in line with the results obtained from the paired t-tests, affirmative judgments were significantly higher than baseline judgments ($p<0.001$) and that rebuttal judgments were significantly higher than affirmative judgments ($p<0.001$). On the other hand, the rebuttal judgments were not significantly lower than baseline judgments ($p=0.09$).

A very similar pattern of results is revealed by running the one-way repeated measures ANOVA on the causal judgments. There was a significant effect for guilt judgment, Wilk's Lambda = 0.624 F (2, 63) = 19.001 $p < 0.001$, multivariate partial eta squared = 0.376. Post hoc tests using the Bonferroni correction revealed that, in line with the results obtained from the paired t-tests, affirmative judgments were significantly higher than baseline judgments ($p<0.001$) and that rebuttal judgments were significantly higher than affirmative judgments ($p<0.001$). On the other hand, the rebuttal judgments were not significantly lower than baseline judgments ($p=0.099$).

Injured child scenario. The same analyses were repeated for the 'Injured child' scenario. A one-way repeated measures ANOVA was conducted to compare the guilt judgments given at the three different stages (after background, after affirmative evidence and after rebuttal evidence). The was a significant effect for guilt judgment, Wilk's Lambda = 0.209, F (2, 63) = 119.22, $p < 0.0001$, multivariate partial

eta squared = 0.791. Post-hoc tests using the Bonferroni correction revealed that, in line with the results obtained from the paired t-tests, affirmative judgments were significantly higher than baseline judgments ($p<0.001$) and that rebuttal judgments were significantly higher than affirmative judgments ($p<0.001$). Additionally, the rebuttal judgments were significantly lower than baseline judgments ($p<0.001$).

A very similar pattern of results is revealed by running the one-way repeated measures ANOVA on the causal judgments. The was a significant effect for guilt judgment, Wilk's Lambda = 0.409 $F(2, 16) = 11.551$ $p < 0.001$, multivariate partial eta squared = 0.591. Post hoc tests using the Bonferroni correction revealed that, in line with the results obtained from the paired t-tests, affirmative judgments were significantly higher than baseline judgments ($p<0.001$) and that rebuttal judgments were significantly higher than affirmative judgments ($p=0.038$). On the other hand, the rebuttal judgments were not significantly lower than baseline judgments ($p=0.202$).

Individual analyses

Individual differences in patterns of responses were analyzed for both sets of judgments. Table 4 shows the number (and percentage) of participants who: i) gave affirmative judgments greater than baseline judgments; ii) gave rebuttal judgments lower than affirmative judgments; gave all judgments consistent with the hypotheses; and iv) participants who did not over-adjust the rebuttal judgment to be lower than baseline.

Table 4.

Table showing individual differences in judgment patterns. Experiment 1.

	House fire scenario		Injured child scenario	
Judgment	Guilt	Causal	Guilt	Causal
Baseline > Affirmative	57 (88%)	49 (75%)	60 (92%)	43 (66%)
Affirmative < Rebuttal	57 (88%)	51 (78%)	60 (92%)	53 (82%)
Baseline > Affirmative < Rebuttal	54 (83%)	42 (65%)	57 (88%)	36 (55%)
Baseline < Rebuttal	23 (35%)	26 (40%)	20 (31%)	28 (43%)

Causal Model assessment

Figure 7 shows the hypothesized causal model for the two scenarios. Table 5 shows the percentage of participants who constructed each link (this was calculated based on the forced choice question. It is the number of participants who responded that a link between the two entities was more likely). The table also shows the mean confidence rating for each participant.

Figure 7. The hypothesized causal model for the two scenarios (C=causal variable; B=guilt variable; E=rebuttal evidence; D=affirmative evidence). The crossed dashed line indicates the absence of the incoherence link. The numbers correspond to the data in the table 5 below.

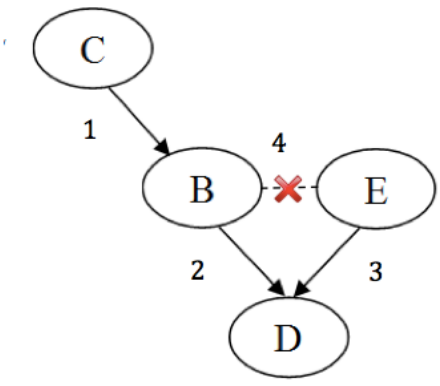


Table 5.

Percentage of participants who constructed each link with the mean confidence rating. A rating of 100 corresponded to 'extremely confident'; a rating of 0 corresponded to 'slightly confident'.

	House fire scenario		Injured child scenario	
Causal link	% constructing link	Mean confidence in link	% constructing link	Mean confidence in link
1	96.9%	46.7 (SD = 30.6)	95.4%	48.5 (SD = 30.8)
2	80%	41.5 (SD = 36.7)	96.9%	47.7 (SD = 30.8)
3	98.5%	75.6 (SD = 31.5)	98.5%	61.9 SD = 29.1)
4	72.3%	11.3 (SD = 19.6)	61.5%	10.7 (SD = 22.9)

The first causal model hypothesis predicted that participants would construct a positive causal link between the guilt variable and the causal variable. This was true for both scenarios. In scenario 1 (House fire) 96.9% judged a person in financial trouble to be more likely to burn down their house than a person not in financial trouble (mean confidence = 46.7, SD = 30.61). In scenario 2 (Injured child) 95.4% judged parents with mental health problems to be more likely to violently shake their child than parents without mental health problems (mean confidence = 48.6, SD = 30.79). The second hypothesis was that participants constructed the guilt variable to be positively linked to the affirmative evidence. This was supported for both scenarios. In the first scenario 80% judged a person who burnt down their house to be more likely to have petrol traces on his clothes than a person who did not (mean confidence = 41.5, SD = 36.71). Similarly, 96.9% judged a child who has been violently shaken to be more likely to have retinal hemorrhage than a child who was not (mean confidence = 47.71, SD = 30.79), in scenario 2. The third hypothesis predicted participants constructed a positive causal link between the rebuttal evidence and the affirmative evidence. Again, this was found to be the case for both scenarios.

In scenario 1, 98.5% judged a person who spilt petrol on his clothes (while filling up his car) to be more likely to have petrol traces on his clothes that a person who didn't spill petrol (mean confidence = 75.6, SD = 31.53). In scenario 2, 98.5% judged a child with vitamin C deficiency to be more likely to have retinal hemorrhage than a child without vitamin C deficiency (mean confidence = 61.9, SD = 29.11). Finally, the last hypothesis was that participants did not construct any casual link between the causal judgment and the rebuttal evidence. This was true for both scenarios. In the first scenario, even though 72.3% judged a person in financial trouble to be more likely to spill petrol on his clothes than a person not in financial trouble, the mean relative likelihood rating was only 11.3 (SD = 19.69). Similarly, for the second scenario, even if 61.5% judged parents with mental health problems to be more likely to have a child with vitamin C deficiency than parents without mental health problems, the mean relative likelihood rating was very low: 10.71 (SD = 22.90).

2.5 Discussion

The aim of the study was to show that mock jurors represent simple crime scenarios in causal networks that allow them to: i) make explaining away inferences, and ii) extend these inferences in line with the causal network. First of all, the experiment showed that when participants were presented with rebuttal evidence that explained away the affirmative evidence, they made explaining away inferences by significantly lowering their guilt judgments. Secondly, this explaining away inference was clearly extended in line with the causal network as participants significantly lowered their causal judgment as well. The combination of these two findings strongly suggests that mock jurors represent problems in causal networks. This was assessed explicitly in the current study and revealed that the links participants had

constructed between the pieces of evidence and their judgments could be represented by the hypothesised causal network.

Importantly, however, it was found that for both scenarios, participants' guilt and causal judgments given after the rebuttal evidence were both lower than the initial baseline judgments. A coherence-based approach could argue that the reason for this is because participants constructed an inhibitory link between the two explanations for the evidence. In theory, this inhibitory link would cause participants to over adjust their guilt and causal judgments below baseline.

Even though this might seem like a compelling argument at first, there is an alternative explanation that might account for these results. It could be that the rebuttal evidence explained away more than just the affirmative evidence. Possibly it explained away some of the background information as well. Participant's guilt judgment represents the probability that the suspect is guilty given the background evidence. If however, the rebuttal evidence explained away some of the background evidence as well as the rebuttal evidence, it is then logical to adjust the guilt judgments below the judgment given at baseline. The same concept follows when it comes to the causal judgment.

2. 6 Experiment 2

Experiment 2 investigated this hypothesis experimentally in a simple study. The experiment replicated the current material but with one important difference: the background information participants were provided with was reduced to a minimum. For instance, taking the first scenario as an example, participants were presented with only the following information: 'A house burnt down. The police investigation reveals that the fire was caused by ignited petrol.' This way, there was less risk that

the rebuttal evidence would explain the background evidence, as this is reduced to a bare minimum. Therefore, it was hypothesized that participants' guilt and causal judgments given after the rebuttal evidence would be the same (or slightly higher) then their initial baseline judgments.

2. 7 Method

Participants and Apparatus

18 first year undergraduate students from UCL (University College London) participated in the study in return for course credit. 12 participants were female and the mean age was 18.9 (SD=1.4). The experiment was conducted online on individual computers and programmed in Adobe Dreamweaver.

Design

All participants were presented with only one scenario (House fire scenario).

Materials and Procedure

The materials consisted of simplified versions of the scenario used in Experiment 1: the 'House fire scenario'. The scenario was accompanied by a set of judgment questions (these were identical to Experiment 1). There were no causal model questions.

The only difference in the scenario was in the background details (the affirmative evidence and the rebuttal evidence were the same). The background of the house fire scenario read as follows: '*A house burnt down. The police investigation reveals that the fire was caused by ignited petrol.'*

2. 8 Results

Scenario judgments

The mean and standard deviation of the three sets of guilt and causal judgments are displayed in Table 6.

Table 6.

Mean and standard deviation of guilt judgments and causal judgments. Experiment 2.

Judgment	Guilt	Causal
Baseline	51.3 (12.1)	50.2 (12.1)
Affirmative	70.3 (11.4)	69 (11.8)
Rebuttal	51.8 (12.9)	57.2 (12.6)

The way the judgments change after the affirmative and rebuttal evidence is clear by observing Figure 8.

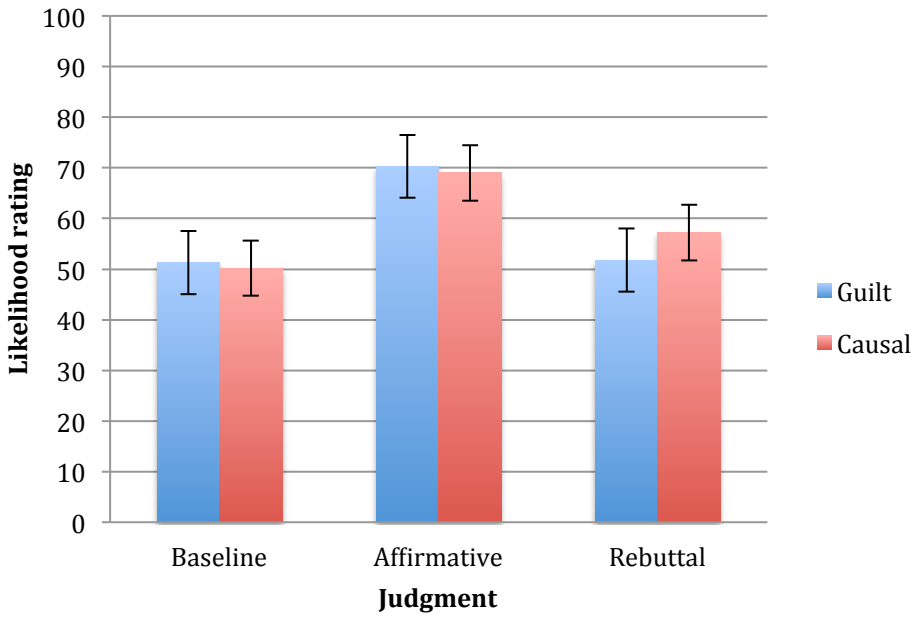


Figure 8. Graph showing the mean guilt and causal judgments (with error bars). Experiment 2.

T-tests

Paired samples t-tests were conducted to evaluate the significance of the results. The first hypothesis was supported as there was a significant increase between guilt judgments given at baseline and those given after the affirmative evidence: $t(17)= -5.127$, $p<0.001$. The same was true for causal judgments (hypothesis 2): $t(17)= -4.878$, $p<0.001$. Also supported was the third hypothesis, which predicted that participants' guilt judgments given after the rebuttal evidence would be significantly lower than the ones given after the affirmative evidence: $t(17)=5.019$, $p<0.001$. The same was true for causal judgments (hypothesis 4) $t(17)=2.783$, $p<0.013$.

A paired sample t-test was conducted to evaluate whether there was a significant difference between the rebuttal guilt judgment and the baseline guilt judgment. The result was not significant, $t(17)=-0.127$, $p=0.901$. The same was true for the difference between the rebuttal causal judgment and the baseline causal judgment, $t(17)=-1.57$, $p=0.135$.

ANOVAs

A one-way repeated measures ANOVA was conducted to compare the guilt judgments given at the three different stages (after background evidence, after affirmative evidence and after rebuttal evidence). There was a significant effect for guilt judgment, Wilk's Lambda = 0.323, $F(2, 16) = 16.8$, $p < 0.0001$, multivariate partial eta squared = 0.677. Post hoc tests using the Bonferroni correction revealed that, in line with the results obtained from the paired t-tests, affirmative judgments were significantly higher than baseline judgments ($p<0.001$) and that rebuttal judgments were significantly higher than affirmative judgments ($p<0.001$). On the

other hand, the rebuttal judgments were not significantly lower than baseline judgments (p=1).

A very similar pattern of results is revealed by running the one-way repeated measures ANOVA on the causal judgments. There was a significant effect for guilt judgment, Wilk's Lambda = 0.624 F (2, 63) = 19.001 $p < 0.001$, multivariate partial eta squared = 0.376. Post hoc tests using the Bonferroni correction revealed that, in line with the results obtained from the paired t-tests, affirmative judgments were significantly higher than baseline judgments ($p<0.001$) and that rebuttal judgments were significantly higher than affirmative judgments ($p<0.001$). On the other hand, the rebuttal judgments were not significantly lower than baseline judgments (p=0.405).

Individual analyses

Individual differences in patterns of responses were analyzed for both sets of judgments. Table 4 shows the number (and percentage) of participants who: i) gave affirmative judgments greater than baseline judgments; ii) gave rebuttal judgments lower than affirmative judgments; gave all judgments consistent with the hypotheses; and iv) participants who did not over-adjust the rebuttal judgment to be lower than baseline.

Table 7.

Table showing individual differences in judgment patterns. Experiment 2.

Judgment	Guilt	Causal
Baseline > Affirmative	16 (89%)	17 (94%)
Affirmative < Rebuttal	16 (89%)	13 (72%)
Baseline > Affirmative < Rebuttal	16 (89%)	13 (72%)
Baseline < Rebuttal	10 (56%)	11 (61%)

2. 9 Discussion

Experiment 1 found that for both scenarios, participants' guilt and causal judgments given after the rebuttal evidence were both lower than the initial baseline judgments. A coherence-based approach could argue that the reason for this is because participants constructed an inhibitory link between the two explanations for the evidence. In theory, this inhibitory link would cause participants to over adjust their guilt and causal judgments below baseline.

Experiment 2 investigated the hypothesis that the rebuttal evidence explained away more than just the affirmative evidence. Therefore the background information participants were provided with was reduced to a minimum (so there was less risk that the rebuttal evidence would explain the background evidence, as this is reduced to a bare minimum). As hypothesized, the simplification of the background information given in the scenario in Experiment 2, yielded a pattern of results that showed no presence of over-adjustment of guilt and causal judgments after the rebuttal evidence.

2.10 General Discussion

The two experiments showed that when participants were presented with rebuttal evidence that explained away the affirmative evidence, they made explaining away inferences by significantly lowering their guilt judgments. Secondly, this explaining away inference was clearly extended in line with the causal network as participants significantly lowered their causal judgment as well. The combination of these two findings strongly suggests that mock jurors represent problems in causal networks. This was assessed explicitly in the current study and revealed that the links participants had constructed between the pieces of evidence and their judgments could be represented by the hypothesised causal network. Importantly, it was found participants did not construct a causal link between the causal variable and the rebuttal evidence.

Implications for the coherence model

This pattern of results cannot be accounted for by coherence-based models. In order for a coherence model to represent the explaining away inference made by participants, it would have to assume that the guilt variable and the rebuttal evidence are exclusive explanations for the affirmative evidence. Returning to the first scenario as an example, a coherence model would have to represent starting a fire and spilling petrol while filling up a car as exclusive explanations for the evidence of petrol traces. In other words, there would have to be an inhibitory link between them.

This is problematic for two main reasons. Firstly, as argued earlier, it creates an unrealistic representation of their true relation in the world. As it is obvious from the example, it is possible that someone spilled petrol while filling up his car and also started a fire – the two explanations are independent but not mutually exclusive. The

fact that these became dependent contingent on the evidence cannot be represented by coherence models, or for that matter, by any current model that relies on bidirectional links to represent relations.

The second reason why constructing an inhibitory link would be problematic is that there is no experimental evidence to support this proposition. The causal model assessment conducted in the current study explicitly investigated whether participants had constructed a link between the two explanations for the evidence. Participants' judgments reflected that they did not construct any link at all between the causal variable and the rebuttal evidence. This provides further endorsement that speaks against coherence based model representation of evidential reasoning.

Implications for the story model

The inappropriateness of coherence-based models for explaining juror reasoning automatically poses a key problem for any evidential reasoning model which integrates the concepts of coherence as one of its key determinants. This becomes an issue of particular importance when it comes to the story model. This is because, as described above, the story model explicitly adopts the coherence concept as a critical tool to resolve which of the stories constructed by jurors should prevail.

This becomes an undeniable problem especially because Pennington and Hastie define coherence as the product of plausibility and consistency. The current results are a direct challenge to the consistency component of the principle of coherence upon which the story model relies on. Pennington and Hastie argue a story is consistent to the extent that it does not contain internal contradictions either with evidence believed to be true or with other parts of the explanation. Secondly, they argue that a story is plausible to the extent that it corresponds to the decision maker's

knowledge about what typically happens in the world and does not contradict that knowledge.

The present study has shown that coherence-based reasoning cannot handle explaining away inferences. This is because it cannot construct a realistic representation of the world and avoid internal contradictions. Naturally this creates somewhat of a catch-22 situation. The story model uses coherence as a criterion based on realistic representations of the world, but the coherence approach does not always represent realistic causal relations in the world. This implies that coherence cannot be used consistently as a valid criterion for deciding which story should prevail. This limitation seriously undermines the explanatory power of the story model as a comprehensive account of juror decision making. This calls either for a reassessment of coherence by the endorsers of the story model or for a different story selection paradigm all together.

Qualitative causal networks and future directions

The shortcomings of coherence-based models, and in turn of the story model, highlight the need to conceptualise a more reliable cognitive model of juror decision making. Lagnado (2011) proposes instead that juror reasoning can be accounted for in terms of qualitative causal networks. He suggests that causal networks are critical in the construction of people's models of the evidence presented in court. This concept has already been emphasised by Pearl (2000) who argued that the best way for people to organise their knowledge of the world is in terms of invariant (stable) qualitative causal relations. One of the key aspects of this proposition is that these relations, once known, will not change according to the particularities of the information at hand, whereas purely probabilistic relations can. For example, returning to the 'explain

away' example cited above, the hypothesis that the owner spilled petrol while filling up his car, and the hypothesis that the owner has petrol traces on clothes because he started a fire, independently explain the evidence and can be represented as: start fire → petrol traces ← spill petrol. On this model 'start fire' and 'spill petrol' are probabilistically independent but become dependent conditional on knowing 'petrol traces'. What remain constant across this change of information in the probabilistic relations (in this instance knowing about 'petrol traces') are the underlying causal relations in the model. This way, by organising knowledge on the basis of invariant rather than unstable aspects of the world, even if causal relations do change, they will reflect a change in the real world rather than change in personal knowledge about it. This aspect of the model discriminates it from the coherence model because the latter treats the two hypotheses as competitors that are incoherent with each other. This, in turn creates an inappropriate representation of their true relation in the world.

Another factor that sets this model apart from the coherence model and the story model, is that the latter two models explicitly reject the idea that people think about evidence in a probabilistic fashion. In contrast, the causal model approach emphasises that people can and indeed do reason probabilistically (as well as causally) but without the need for precise numerical estimates. For example the presence of petrol traces on a suspect's clothes increases the likelihood of guilt, even if exact probabilities cannot be assigned. Lagnado (2011) proposes that such important patterns of inference can be represented using a formal Bayesian framework. Following this argument, ongoing research is assessing to what extent people's qualitative causal networks reflect an underlying Bayesian reasoning process (Lagnado, Fenton and Neil, 2013; Fenton, Neil and Lagnado, 2013).

Another shortcoming with current legal decision making research is that juror reasoning is often treated as an isolated process, overlooking how it is modulated by other cognitive resources. In order for the proposed qualitative causal network model to achieve robustness, it will be fundamental to consider how jurors' representations and thinking is modulated by other concurrent reasoning processes and working memory. Future research exploring these factors will help pin down the key structural features of jurors' representations, such as the size of network they can reason with. Furthermore, it will shed light on crucial reasoning processes such as how jurors gain access to and update the fragments of the networks they construct. To conclude, this research evolving around the causal network approach seems a promising starting point in achieving a valid understanding of the psychological processes underlying juror decision making: a step of fundamental importance to the criminal justice system.

Chapter 3: Reasoning with causal networks: understanding environmental problems

3.1 Introduction

Background

Despite its crucial importance for the survival of humanity, now more than ever before, marine biodiversity is at the brink of collapse. After climate change, the main culprit is overfishing. This occurs when fish and other marine species are caught faster than they can reproduce. According to the Food and Agriculture organization (1995), over 70 percent of the world's fisheries are "fully exploited", "over exploited" or "significantly depleted". In fact, some species have already been fished to commercial extinction, and more are on the verge of disappearance. Simply put, overfishing is the outcome consumers' growing demand for seafood around the world, combined with poor management of fisheries and development of new, more effective fishing techniques. Unfortunately, the devastating effects of overfishing are not limited to loss of biodiversity and ecosystems. One billion people rely on fish as a key source of daily protein, and as the main source of food in many developing countries. Millions of people, and entire coastal communities, depend on fisheries for their employment. In addition, and often forgotten, collapsing fish stocks aggravate climate change by creating large ecological dead zones with no oxygen in the oceans.

Despite these tragic consequences, overfishing, like most environmental problems of this day and age, is another complex tragedy of the commons. The

extensive number of causes that surround the problem generates the element of complexity. Furthermore the interactions between these causes tend to unfold across different time frames and inevitably become too convoluted for their effects to be predictable. Nonetheless, consumers, or lay people, still manage to form beliefs about the roles of these various causes and their probable effects. In light of the complexity involved these beliefs are likely to be oversimplified and inaccurate (White, 2008). As oversimplified or partial as they may be, these beliefs may still form the basis of ordinary people's understanding and therefore reasoning about the phenomenon.

Causality in environmental decision-making

White (2008) pointed out that since ordinary people can make a difference by the force of their opinions and values, there is an obvious practical importance in ascertaining the content and structure of lay beliefs about causal processes related to environmental problems. When it comes to overfishing, even if most of the damage cannot be reversed, a lot can be saved and eventually restored if people started making sustainable decisions. Therefore understanding how people's causal representation of overfishing may mediate the way they reason about the problem has significant practical implications.

The value of discovering naïve causal structures of a given environmental problem resides in the fact that they convey how people understand the problem as a system of interconnected parts (White, 2008). In other words, individual causal relations are not isolated from each other but tend to be linked in dynamic systems. Given that environmental problems are the product of an entrenched system made up of numerous interdependent actors, they tend to be immensely intricate. For this reason, it does not make sense, nor it is fruitful, to focus on isolated causal beliefs.

Instead, as White (2000) points out, the scope should be to integrate 'collections of individual events into an organized representation of chains and networks of causal relations'.

Past research in causal cognition has already explored complex causal structures (Lagnado, Waldmann, Hagmayer, & Sloman, 2007; Sloman 2005; Waldmann, Hagmayer & Blaisdell, 2006). A central finding from this research is that people do indeed create and use causal models to structure their learning and inference. However, only a few studies have looked at the role of feedback loops (e.g. Kim, Luhmann & Ryan, 2009) or dynamical systems (e.g. Rottman & Keil, 2011).

Consequently, especially when it comes to environmental problems, not much is known about how individual causal beliefs are related in an overall dynamic causal structure and the form of that structure. For example, knowing that people believe that destructive fishing gear causes overfishing carries little meaning unless this individual belief can be placed in a network of beliefs about entities related to fishing gear and overfishing. People might also believe that their consumption of unsustainable fish is not related to the use of destructive fishing gear. If this is their causal model of overfishing, then people may believe that consumption of unsustainable fish does not affect overfishing at all. In other words, naïve causal models need to be ascertained as a network of causal beliefs.

Causal network analysis

Methods to examine the relationship between someone's causal model and their actions remain underdeveloped. However, the most promising and established

method to achieve this is causal network analysis (Green & McManus, 2003). Network analysis of causal beliefs is a method that has been pioneered by Lunt (1988) in a study of perceived causes of failure. Essentially, causal network analysis is a method of representing a set of causes and links between them. For example, in Lunt (1988) the entities were possible causes of failure, such as having little intelligence or poor concentration, and the links were judged causal relations. Lunt's study revealed a network of interrelations between these different causes. For instance, low intelligence was causally related to poor concentration, and both were deemed to lead to poor time allotment.

Inspired by this methodology, Green and McManus (1995) explored the idea that individuals construct causal models of reality and use them to think about possible actions in the world. Green and McManus (1995) employed the causal network analysis method to examine individuals' causal model of risk factors for coronary heart disease (CHD) and related these to their judgments of preventive actions. They required individuals to draw a network diagram of the risk factors, or causes, for CHD (e.g. high blood pressure, fatty foods and exercise). Specifically, individuals were asked to represent a causal relationship between two factors by drawing a line connecting them. They were also asked to indicate the direction of the causal influence using an arrowhead. In the diagrams created, a causal factor could be connected to another factor in a variety of ways. It could have a direct path to the connected factor, or it could have an indirect path to it via some other factor, or it could have both a direct path and an indirect path to the target factor. For example, they found that eating fatty foods was deemed to increase the risk of CHD directly, but also indirectly through increasing cholesterol. Therefore the diagram represented what individuals spontaneously considered to be the critical pathways. In addition to

representing a path, individuals were required to rate the strength of each causal path on a scale from zero to one hundred. The same individuals also rated the effectiveness of different preventive actions related to the factors (e.g. reducing blood pressure).

Green and McManus (1995) showed that the total path strength of a factor (the strength of both direct paths and all indirect paths) predicted participants' ratings of the effectiveness of the different preventive actions. Total path strength accounted for two thirds of the variance in these ratings. In a subsequent study, Green, McManus and Derrick (1998) examined perceptions of a person's prospects of employment. They confirmed the importance of path strengths in predicting the ratings of effectiveness of different actions designed to increase a person's employment prospects. Furthermore, it extended the previous findings by showing that the total path strengths added considerably more than just the direct paths; almost doubling the variance explained in the effectiveness ratings.

Feedback loops

Green and McManus' analysis involved investigating the presence or absence of causal paths as well as the strengths of those paths. However, another advantage of causal network analysis is that apart from allowing investigation of its content features, it allows inspection of its structural features. White (2008) argued that causal models could have various kinds of embedded structures that convey broad ideas about how people understand the phenomenon (e.g. overfishing). Based on this idea, he conducted a series of studies concerned with the structure of people's beliefs about causal processes in complex natural environments. His main aim was to discover whether people construct causal processes in nature in a systems-like

manner, involving one-way causal hierarchies. White (2008) derived a set of factors in relation to forest ecosystems and climate change (e.g. human population, atmospheric carbon dioxide levels, fires and several biological features such as extinction rates). In two experiments participants were presented with each pair of factors and asked whether change in one would produce change in the other. From the participants' judgments, he constructed a causal network that reflected consensual causal beliefs. This method of network elicitation has been termed the "grid method" (Green et al. 2003) and differs from the one employed by Green et al. (1995, 1998).

White found the resultant causal network to be unidirectional. In other words, the network did not encompass any feedback loops, but was composed of linear causal chains mostly arranged in a unidirectional hierarchy. Some factors, such as humans, functioned as causal origins and others, such as extinction rates, functioned as effects. These results are consistent with White's previous research (1992a, 1995a, 1997, 1999 and 2008) showing unidirectional patterns of thinking about causality in natural systems. White argues that such findings reveal a general failure to appreciate the interactive processes that govern the operations of natural systems.

Conversely, the results do not mean that people are unable to create interactive models of ecological systems. Green (2001) presented participants with a food web and asked participants to explain a complex pattern of fluctuation over time in the population of an herbivore. He found that most people were able to construct interactive accounts involving two, and in some cases three, entities (plant, herbivore and carnivore). However, as argued by White (2008), the system, or food web, comprised only three entities and individuals were constrained to explain a complex pattern presented to them, rather than envisaging themselves what sort of pattern might occur. It is therefore not clear whether the interactive thinking exhibited in that

study is characteristic of reasoning about ecological systems outside the psychological laboratory. Nonetheless, studies by Green (1997; 2001) do suggest that people have a capacity to think about interactions in natural systems, but that this capacity might be overwhelmed by task complexity.

The complexities of interactive systems with multiple entities are admittedly hard to grasp, but failure to fully appreciate interactions and feedback loops in these systems could have detrimental consequences for the global ecosystem. In other words, how humans treat the world must to some extent reflect what they believe about the effects of that treatment - if people believe that anything can be done to nature, without repercussions for the human world, they are less likely to exhibit sustainable behavior. Kempton (1986) pointed out that lay models about physical systems influence real life decision-making. He found that people's mental models of thermostats accounted for how they treat the control of heat in their homes. Those who possessed one theory tended to behave more economically than those who possessed the other theory. Kempton (1986) proposed, on the basis of interviews, that people used two distinct models of home heating systems. In the (incorrect) valve model, the thermostat is thought to regulate the rate at which the furnace produces heat. Therefore setting higher makes the furnace work harder. In the (correct) threshold model, the thermostat is viewed as setting the goal temperature, but not as controlling the rate of heating. Hence the furnace runs at a constant rate. Kempton then examined thermostat records from real households and found that the patterns of thermostat settings fitted nicely with the two models he had found.

As another example, Atran, Medin, and Ross (2005) found that cultural groups' mental models of plant/animal interactions in the rainforest were consistent with the environmental impact of those groups. Therefore, common-sense

understanding of the structure of cause and effect in nature, with specific focus on understanding of cyclical interactions between humans and nature, is an important topic from both a scientific and conservationist point of view.

3.2 .Experiment 1

The main aim of the current research is to investigate how the lay representation of the causes of overfishing may underlie individuals' reasoning about the problem. The first aim of Experiment 1 is to elicit individual causal models of overfishing in order to examine the relationship between these models and the ratings of effectiveness of various related actions.

Green et al. (1995, 1998) elicited individuals' causal models through the causal network diagram task and showed that total path strength of a causal factor predicted participants' ratings of the effectiveness of different actions related to the diagrammed factors. These results were found in two naturalistic domains: representations of risk factors related to CHD and causes of unemployment. The main aim of Experiment 1 is to investigate whether this connection between a casual model and the assessment of actions can be extended to the environmental problem of overfishing. In other words, this study seeks to determine whether people's individual causal representation of overfishing predicts the way they reason about the issue.

Experiment 1 employs the method advocated by Green et al. (1995; 1998) to elicit participants' causal network of overfishing. One potential weakness of the causal network analysis is that the network obtained, and its structural features, depend on the factors selected for the study. Clearly, there are many possible factors one could include as causes of overfishing (and each of these could be unpacked

almost ad infinitum). As White (2008) points out, one solution is to rely on expert assessments of the relative importance of different factors. To this end, a number of expert sources (e.g. Hilborn, 2012) were reviewed and five factors were selected as the main causes of overfishing: consumption of unsustainable fish, poor monitoring and enforcement of fishing laws, fishermen using gear that does not permit capturing only the targeted species, unsustainable fish being sold on the market and demand for unsustainable fish.

Participants were also asked to evaluate the effectiveness of different actions based on these diagrammed factors. These questions were phrased as counterfactual questions and the response was in the form of a quantitative judgment. Subjects are told to imagine that a 30% change (increase or decrease) has occurred in the factor in question and are asked to judge the amount of change this would cause to overfishing. So, for example, subjects are told to imagine that there has been a 30% increase in consumption of unsustainable fish. They are then asked to say whether there would be an increase, decrease or no change in overfishing. For the former two, they are asked to give an estimate of the amount of change that would occur in percentage.

The counterfactual questions encourage participants to reason about the counterfactual suppositions as if they were external interventions on overfishing. Sloman and Lagnado (2005) showed that when reasoning about the consequences of a counterfactual supposition of an event, most people do not change their beliefs about the state of the normal causes of the event. Therefore, when participants answer these questions they should not change their causal beliefs about overfishing, but just reflect upon the effect of the mentally changed event (e.g. consumption). In addition, the counterfactual supposition involved a quantified change (30%) to ensure all participants simulated the same amount of change. This also means that judgments

across the different questions were more comparable to each other. Hence, any relationship between the diagrams and the counterfactual judgments should be revealed by a positive correlation between total path strength for each factor (direct and indirect paths) and the judgments.

In addition to computing total path strength, Green and McManus (1997) also computed the direct path strengths alone, and the total number of paths emanating from each factor (ignoring their strength). The same analysis will be carried out in the present study. If the perceived strength of a causal path is important, then total path strength should correlate significantly more highly with the counterfactual judgments than direct path strengths alone or number of paths.

The second aim of Experiment 1 is to investigate whether people think about overfishing in an oversimplified linear and unidirectional way. Previous studies by White (1992a, 1995a, 1997, 1999, 2008) investigated the structural features of causal networks, namely the presence of feedback loops, on the consensual network. In other words, he constructed the causal model based on the aggregated data from all participants. This method has two main drawbacks. First, there are obvious theoretical problems in deciding on an appropriate threshold for the inclusion of paths. There are different thresholds that can be used and these yield very different causal networks that vary in the degree of complexity and therefore structure. Second, even though the consensual representation of a phenomenon may be interesting in its own right, there is a lot of variety and individual differences in the individual causal representations. These could involve significant structural features, such as feedback loops, that get obscured in the creation of the consensual model because they might differ in the type, or number of factors they comprise.

For this reason, Experiment 1 will adopt a novel approach and will investigate

the presence of feedback loops at the individual level. The causal network diagram method of elicitation encourages participants to focus on the overall structure of the network, which is visible to the participant as they proceed with the line-drawing task. Thus participants never lose sight of the overall structure of their beliefs. Consequently, any feedback loops drawn are likely to reflect genuine causal beliefs. In the study by Green et al. (2003), the diagram method yielded a network with no feedback loops, so there is no evidence that a graphical method improves the likelihood of obtaining feedback loops. Therefore, employing this methodology might also provide a more rigid test of representation of feedback loops.

3.3 Method

Participants

40 participants were recruited through the University College London Psychology Subject Pool. The subject pool in question is open to everybody and therefore not limited to university students. The study was advertised as investigating reasoning about causes and effects. All participants were paid £4. Twenty-four participants were males (60%) and 16 were females (40%). The mean age was 30.9 years (SD = 14.1; range 19 to72 years). Thirty participants completed the task satisfactorily; the remaining 10 either failed to label all paths with an indication of direction or failed to give a numerical estimate of strength for each of the paths. The participants' environmental values were measured through the New Environmental Paradigm (NEP) scale (further details discusses in *Materials* section). They were a representative sample of the general population in terms of environmental values (mean NEP score was 21.7, SD=3.63).

Design

The order in which participants completed the diagram task and the counterfactual judgment task was counterbalanced. The order in which the casual factors were presented in the diagram task and the order in which the counterfactual judgments questions were presented, were both randomized. The framing of the counterfactual judgments (increase or decrease frame) was counterbalanced. The questionnaire ended with a series of demographic questions.

Materials

The materials consisted of a written questionnaire. The first page of the questionnaire provided a simple definition of overfishing followed by a few sentences detailing some of its effects (e.g. environmental problems). In addition, participants were informed that the survey was part of a project to discover the best approaches to decrease overfishing. The second page was an instruction sheet. Then, depending on the counterbalancing condition, participants were either given the diagram task followed by the counterfactual judgments task, or vice-versa. Following both tasks, participants were given a series of demographic questions including the New Environmental Paradigm. Pro-environmental values were measured using a reduced (6-item) version of the New Environmental Paradigm (NEP) scale (alpha=0.7) (Dunlap, Van Liere, Mertig, & Jones, 2000).

The causal diagram task. Participants were asked to draw a diagram indicating how, in their view, a set of causes or factors are linked to overfishing and to each other. They were instructed as follows:

> There are a number of causes or factors as explanations of overfishing. We would like you to draw a diagram (on the next page) of how you think various factors (listed below) are linked to overfishing and each other, using arrows to

indicate the direction of the effect. Label each arrow with either "increases" or "decreases" to clarify the type of effect. The names of the factors to be diagramed are listed below. Beside each factor you will find a short description. The expression "unsustainable fish" will be used throughout the survey. This simply means fish that are overfished or caught or farmed in ways that harm other marine life or the environment.

The 5 factors that were presented to participants are reported in Table 1. The order in which they were presented was randomized. *Overfishing* was also included in the list.

Table 1.

Causal factors presented in the diagram task, Experiment 1.

Factor names	Interpretation
Consumption	People buy and consume unsustainable fish.
Demand	There is demand for unsustainable fish.
Market	Unsustainable fish is sold on the market.
Overfishing	Fishermen catch unsustainable fish (resulting in overfishing).
Monitoring	The government monitors and enforces fishing laws.
Gear	Fishermen use fishing gear that does not permit capturing only the targeted species.

Participants were told to indicate the connection of these factors (which could be either direct or indirect) to overfishing by including *overfishing* in their diagram. Finally, participants were presented with an example of a schematic diagram, which bore no factor names, as an example.

After drawing the diagram, participants were instructed to rate the strength of

each of the links they drew on the previous page. They were told go back to the previous page and write a number between 0 and 100 where 0 meant no relation and 100 meant an invariable relation. An example was provided to clarify (inspired by Green et al., 1998): "So, for instance, when water boils at 100 degrees °C, steam comes off. There is an invariable relation between the two."

The counterfactual judgment task. The instructions for the counterfactual judgment task were as follows:

> There are a number of factors that may affect overfishing. We would like you to evaluate how certain changes in certain factors may affect the amount of overfishing (the extent to which fishermen catch unsustainable fish). The amount of change is always given as 30%: this is just a convenient figure with no special significance. Your task is to decide whether the change will cause an 'increase', 'decrease' or 'no change' in the amount of overfishing. When you have decided, put a circle round the answer you've chosen. If you've chosen increase or decrease, please also write your estimate of how much change will occur in the space provided. You should do this by giving a percentage estimate, from 1 to 100 per cent (0 per cent would be no change). If you choose 'no change' you do not need to give a percentage estimate. It isn't easy giving an exact percentage judgment, but please do the best you can, basing your judgments on your understanding of how things work. This not a test and there are no right or wrong answers, we are simply interested in the way people think about these things.

Participants were then presented with four questions. Subjects are told to imagine that a 30% change (increase or decrease) has occurred on the factor in question and are asked to judge the amount of change this would cause on overfishing. So, for

example, subjects are told to imagine that there has been a 30% increase in consumption of unsustainable fish. They are then asked to say whether there would be an increase, decrease or no change in overfishing. For the former two, they are asked to give an estimate of the amount of change that would occur in percentage. There was no question related to the factor *demand.* The question about *demand* would have been very similar to the one about *consumption* and could have confused the participants. An example of a question, related to *consumption* is shown below:

> Imagine that people decrease their consumption (eating and buying) of unsustainable fish by 30%.
>
> What effect would this have on overfishing (the extent to which fishermen catch unsustainable fish)?
>
> ☐ Would increase overfishing.
>
> ☐ Would decrease overfishing.
>
> ☐ Would cause no change in overfishing.
>
> How much change would occur? Please write a number from 0 to 100%.
>
> ___%.

Procedure

Participants took part individually or in groups of two or three in a large seminar room. If in groups, participants were positioned so that nobody could see what the others were doing. Participants were supervised by an experimenter who introduced the study, handed out informed consent forms and invited participants to ask questions if anything in the instructions was not clear. There were no questions concerning the present study. At the end, participants were thanked and given their

pay as well as a debriefing sheet that explained the aims of the research. The whole study lasted on average 25 minutes.

3.4 Results

Causal network analysis

The 30 participants included an average of 7.9 paths in their diagrams (SD=2.28; range 4 to 12). For the causal network analysis the quantitative estimates were disregarded, and only the direction of change (increase, decrease or no change) was considered (White, 2008). The data from each subject therefore consisted of a 6 x 6 matrix. Any given cell in the matrix could contain either + (judged increase), - (judged decrease) or 0 (judged no change). Scoring judged increase as +1 and judged decrease as -1 enables a net score to be calculated across subjects for each cell of the matrix. For example, for the cell representing *market* as the cause and *overfishing* as the effect, 6 subjects judged an increase, 1 judged a decrease, and the remainder (24) judged no change. This yields a net score of 5 (6 -1). This means that the causal relation *demand-overfishing* has a net score of five. This matrix is presented in Table 2. These net scores formed the basis for the construction of the causal network.

Lunt (1988) proposed two criteria for selecting links to be included in the causal network. One is the "minimum systems criterion" (MSC). This is the value at which all causes are included in the system, to determine the network nodes. Accordingly, causal links are added hierarchically to the network, in order of net scores, until the MSC is reached. In this study the MSC was a net score of 11 (as used by White, 2008), with 9 links meeting this criterion. Each link in this network was endorsed by at least 36.7% of participants, suggesting a low consensus amongst the participants. The resultant network is shown in Figure 1

Table 2.

Endorsement frequencies of causal links, Experiment 1.

	Effect					
Cause	1	2	3	4	5	6
1 Demand	-	6	14	1	2	16
2 Consumption	15	-	15	0	3	15
3 Market	5	14	-	-1	2	14
4 Monitoring	-2	-3	-6	-	-6	-16
5 Gear	1	0	1	0	-	11
6 Overfishing	2	5	4	2	1	-

The second criterion for selecting links to be included in the causal network is Inductive Eliminative Analysis (IEA), wherein every network produced when working towards the MSC is checked for endorsement. Originally developed to deal with binary adjacency matrices, networks were deemed consensual if endorsed by at least 50% of participants (Brogan & Hevey, 2010). The resultant network is shown in Figure 2.

Figure 1. *Consensual Causal Network created using MSC, Experiment 1. Dashed line indicates a causal relation labeled as "decrease".*

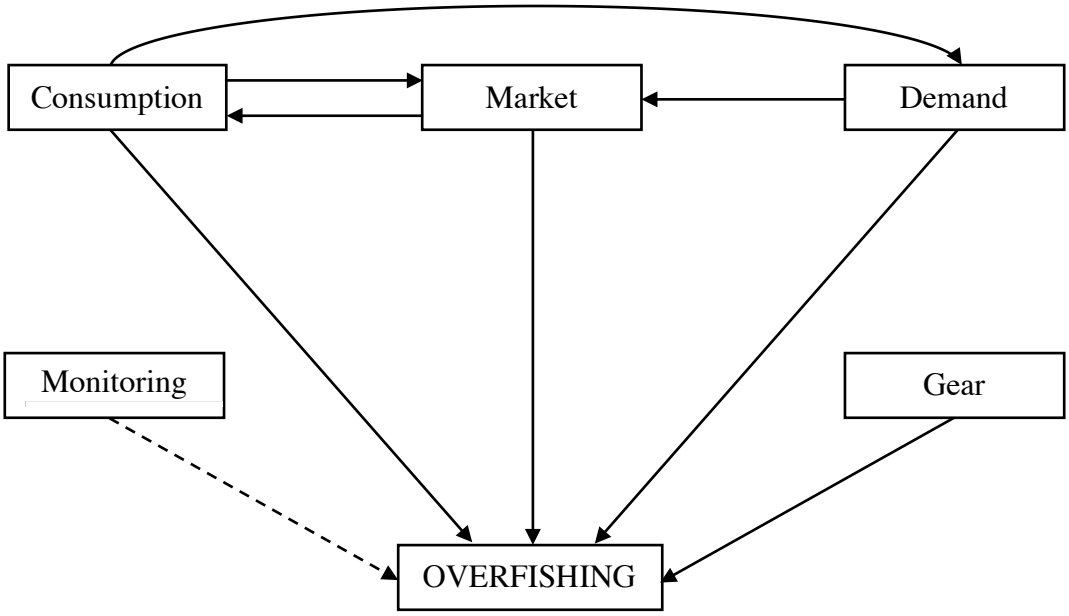
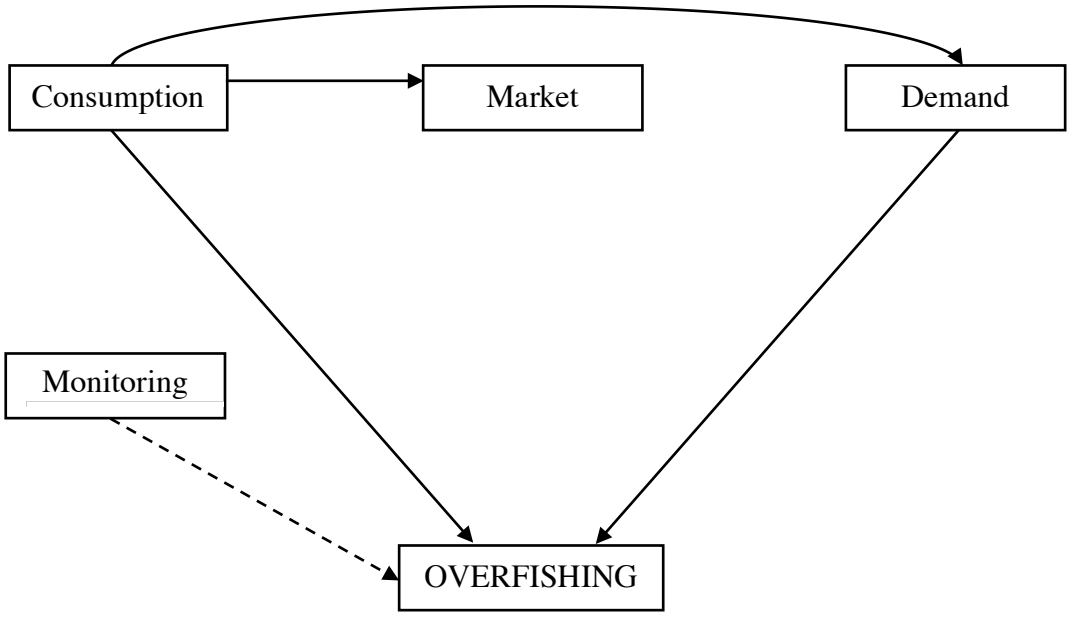


Figure 2. *Consensual Causal Network created using IEA, Experiment 1. Dashed line indicates a causal relation labeled as "decrease".*



.

In this instance adopting IEA resulted in a more conservative network with only 5 links and no feedback loops. As the level of endorsement of the MSC is very low (36.7%), it is likely that in this case IEA yielded a more realistic representation of consensual beliefs. The number of participants representing each path in the resulting consensual networks, as well as the judged mean strength, is reported in Table 3.

Table 3.

Paths and mean path strengths of consensual network created using MSC, Experiment 1.

Cause factor	Effect factor	N (%) representing the path	Mean path strength (SD)
Demand	Overfishing	16 (53.3%)	0.62 (0.27)
Monitoring	Overfishing	16 (53.3%)	-0.62 (0.27)
Consumption	Overfishing	15 (50%)	0.8 (0.29)
Consumption	Market	15 (50%)	0.67 (0.31)
Market	Overfishing	14 (46.7%)	0.85(0.22)
Demand	Market	14 (46.7%)	0.87(0.16)
Market	Consumption	14 (46.7%)	0.76 (0.27)
Gear	Overfishing	11 (36.7%)	0.61 (0.25)

Causal models and counterfactual judgments

The first research question concerned the relationship between a person's causal diagram of overfishing and their counterfactual judgments. In order to examine the relationship between diagrams and judgments, Green et al. (1995; 1998)

used a method that is formally identical to path analysis. Participants' ratings of path strength are treated as being equivalent to standardized path coefficients. The same method is used to address the current question.

First, the total path strength of each factor to overfishing was calculated. For example, there might be a direct path from factor A to overfishing and an indirect path via factor B. In that case, the total path strength from factor A to overfishing would be an additive combination of the strength of the direct path, and the strength of the indirect path. The strength of the indirect path is the strength of the path from A to B (e.g. 20 per cent or 0.2) times the strength of the path from B to the target (e.g. 30 per cent or 0.3). In this instance it would be 0.06 (0.2 times 0.3). If the strength of the direct path from A to the overfishing were 0.4 (40 per cent or 0.4) then the total path strength would be 0.46 (0.4+0.06) and so on for any more complex set of paths between any two factors. The mean path strengths are reported in Table 4. Table 5 shows the mean (SD) counterfactual judgment for each factor (the factor *demand,* as noted in the Method section*,* was excluded as it confounded with *consumption*).

The second step involved calculating, separately for each individual, the correlation between each of the factors' total path strength and those same factors' corresponding counterfactual judgment, using a conventional Pearson correlation r. In the present experiment, the mean correlation across individuals (i.e. the mean of a set of r correlations), was 0.66 (SD = 0.36, N = 30), accounting for 44% of the variance in the counterfactual judgments. This correlation was significantly different from zero, $p<0.001$. This correlation was also higher than the correlation between these counterfactual judgments and the direct path strengths alone (0.63, SD = 0.4). However this difference was not significant: $t(29)= .5$, $p=0.617$. The correlation between judgments and indirect path strengths alone was 0.47 (SD=0.45). The

correlation between judgments and number of paths emanating from each factor was -0.55 (SD=0.31).

Task order did not have an effect on any of the variables. NEP scores did not vary as a function of individual causal models or counterfactual judgments.

Table 4.

Mean (SD) total path strengths from each factor to overfishing, Experiment 1.

Factor	Mean (SD)
Demand	0.83 (0.71)
Consumption	0.98 (0.9)
Market	0.79 (0.69)
Monitoring	-0.71(0.77)
Gear	0.22 (0.52)

Table 5.

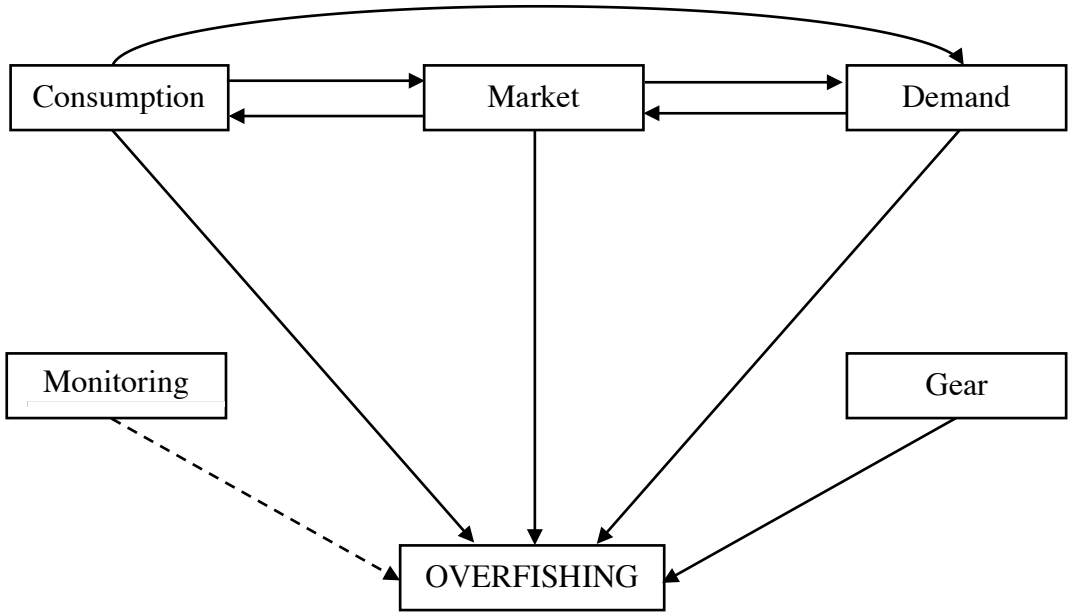
Mean (SD) counterfactual judgments, Experiment 1. Judgments based on the decrease frame have been reversed to compute the mean.

Factor	Mean (SD)
Consumption	28.8% (26.9)
Market	26.5% (24.8)
Monitoring	26.4% (16.3)
Gear	7.8% (35.7)

Analysis of feedback loops

The causal network analysis reported above suggests low consensus amongst the participants and makes an even stronger case for analyzing networks on an individual basis. Each individual causal network was inspected for the presence of feedback loops. 59% of networks had at least one feedback loop. An example of an individual network is show in Figure 3. The mean number of loops across all networks was 1.7 (SD=2.2). The mean number of loops across only the networks containing loops was 2.72 (SD=2.24), ranging from 1 to 8 loops per network. Loops involved 2 to 5 factors. The most common loop (N=9) was *market-consumption*.

Figure 3. *Example of an individual network with feedback loops, Experiment 1.*



3.5 Experiment 2

In line with the experimental hypothesis, the total path strength of the diagrammed factors correlated more significantly with the counterfactual judgments

than direct path strengths alone, or simply the number of paths emanating from a factor. In addition, analysis of individual networks showed that over half of the participants built feedback loops into their network diagrams.

Experiment 1 asked participants to diagram five factors, but answer counterfactual judgments based on only four of the 5 factors. It was not possible to discard the fifth factor (demand) from analysis of the causal networks. Therefore computation of the total path strengths had to encompass links to and from the fifth diagrammed factor (demand). Consequently, the first aim of Experiment 2 is to replicate the method and findings of Experiment 1 but excluding *demand* as one of the factors participants are asked to diagram.

The second aim of Experiment 2 is to investigate the nature of the individual differences in the cognitive mechanisms that might underlie construction of feedback loops. Kim et al. (2009) noted that a true representation of a feedback loop (A causes B, B causes A) is unrealistic, in that loop features often do not cause each other constantly and simultaneously but, rather, unfold over time. Instead, feedback loops might be morerealistically represented as causal chains that play out over time. Such a chain would allow factor A to influence factor B at Time 1, factor B to influence factor A at Time 2, and so on.

Therefore, it seems plausible that a construct such as the ability and propensity to think about the future might be associated with the representation of feedback loops. Such a construct is termed "consideration of future consequences" or CFC (Stratham, Gleicher, Boninger, & Edwards, 1994). CFC measures individual differences in thinking about the future implications and outcomes of one's behaviour and the extent to which these outcomes influence behaviour (Stratham et al., 1994). The CFC scale comprises 12 questions and had shown good internal consistency, with

reported Cronbach alphas ranging from .80 to .85 (Stratham et al., 1994). Previous studies have associated high CFC scores with a range of environmental behaviors. Examples include preference for public transport and beliefs about the negative environmental impact of cars (Joireman, Lange & Van Vugt, 2004), as well as support for public transportation plans which were perceived as reducing pollution (Joireman, Van Lange, Van Vugt, Wood, Vander Leest & Lambert, 2001). Individuals scoring high on the CFC scale have also expressed less support for offshore oil drilling (Stratham et al., 1994) and have been more likely to buy energy-efficient light bulbs when the savings associated with the purchase were framed in the future and not on immediate gains (Tangari & Smith, 2012). Finally, stronger pro-environmental attitudes, intentions and behaviour (Milfont & Gouveia, 2006; Joireman, Lasane, Bennett, Richards & Solaimani, 2001), personal optimism (O'Brien & Brittain, 2009), greater conscientiousness (Stratham et al., 1994), and cooperation in resource dilemmas (Joireman, Posey, Truelove & Parks, 2009; Kortenkamp & Moore, 2006) have all been linked to future orientation, as measured by the CFC construct. Experiment 2 will measure propensity to think about the future through the CFC scale and will investigate the relation between CFC scores and construction of feedback loops.

3.6 Method

Participants

40 participants were recruited through the University College London Psychology Subject Pool. The subject pool in question is open to everybody and therefore not limited to university students. The study was advertised as investigating reasoning about causes and effects. All participants were paid £4. Eleven of them

were males (27.2%) and 29 were females (72.5%). The mean age was 23 years (SD = 4.18; range 17 to 37 years). Thirty-three participants completed the task satisfactorily; the remaining 7 either failed to label all paths with an indication of direction or failed to give a numerical estimate of strength for each of the paths. The participants were a representative sample of the general population in terms of environmental values (mean NEP score was 22.12, SD=3.75).

Design

The order in which participants completed the diagram task, the counterfactual judgment task and the CFC scale, was counterbalanced. The order in which the casual factors were presented in the diagram task and the order in which the counterfactual judgments questions were presented, were both randomized. The framing of the counterfactual judgments (increase or decrease frame) was counterbalanced. The questionnaire ended with a series of demographic questions.

Materials

The materials consisted of a written questionnaire. The first page of the questionnaire provided a simple definition of overfishing followed by a few sentences detailing some of its effects (e.g. environmental problems). In addition, participants were informed that the survey was part of a project to discover the best approaches to decrease overfishing. The second page was an instruction sheet. Then, according to the counterbalancing condition, participants were either given the diagram task followed by the counterfactual judgments task, or vice-versa. Following both tasks, participants were given two questions concerning their fish consumption and finally a series of demographic questions.

The causal diagram task. The instructions and materials related to the causal diagram task were suitably modified from Experiment 1. The factor *demand* was

excluded. Therefore participants had to diagram only the four remaining factors: *consumption, monitoring, gear* and *market*. Finally, rather than giving participants an example of a schematic diagram with no factor names, a concrete example was presented. This was to see if a clearer example could reduce the number of participants who do not complete the task. The example was related to a causal diagram of factors related to failing a Math exam. After drawing the diagram, participants were instructed to rate the strength of each of the links they drew on the previous page in the same way as in Experiment 1.

The counterfactual judgment task. The counterfactual judgment task was identical to Experiment 1.

The CFC scale. Consideration of future consequences was assessed by means of the 14-item CFC measure reported by Joireman et al. (2012). CFC-14 comprises 14 items, half of which assess concern with future consequences (e.g. "I think it is important to take warnings about negative outcomes seriously even if the negative outcome will not occur for years") and half with immediate consequences (e.g. "I generally ignore warnings about possible future problems because I think the problems will be resolved before they reach crisis level"). Respondents were required to indicate to what extent each item characterized them on a 7-point Likert-type scale (1 = very uncharacteristic of me; 7 = very characteristic of me). Higher scores on CFC-future indicate more consideration of future consequences, whereas higher scores on CFC-immediate indicate more consideration of immediate consequences. Both the Future and Immediate factors have previously shown good internal reliability (Cronbach's αs = .80 and .84 respectively). Joireman, Shaffer, Balliet, and Strathman (2012) advocated the use of the 14-item scale version rather than the original 12-item version (Strathman et al., 1994) because it provides two balanced

CFC subscales (both 7-item subscales) and improves upon the internal reliability of the original 5-item CFC-Future subscale (Joireman et al., 2008).

Consumption questions. Participants were asked to indicate how many times a month they consumed fish. In addition they were asked whether they took sustainability or overfishing into account when purchasing fish.

Procedure

The procedure was as in Experiment 1.

3.7 Results

Causal network analysis

The 33 participants included an average of 8 paths in their diagrams (SD=3.3; range 3 to 20). The full matrix of endorsement frequencies is presented in Table 6. These are net frequencies, obtained by subtracting the number of decrease judgments from the number of increase judgments. Positive numbers indicate net judged increases and negative numbers indicate net judged decreases.

Table 6. *Endorsement frequencies of causal links, Experiment 2.*

	Effect				
Cause	1	2	3	4	5
1 Consumption	-	21	1	7	20
2 Market	20	-	4	8	9
3 Monitoring	-7	-22	-	-15	-24
4 Gear	1	3	1	-	21
5 Overfishing	20	2	6	0	-

In this experiment the MSC was a net score of 21, with 4 links meeting this criterion. The resulting network is shown in Figure 4. Each link in this network was endorsed by at least 63.7% of participants, suggesting a considerably higher consensus than in Experiment 1.

As the consensus for the MSC is above 50%, applying the IEA criterion resulted in a more stringent network with 2 less links. This network is shown in Figure 5.

Figure 4. *Consensual Causal Network created using MSC, Experiment 2. Dashed line indicates a causal relation labeled as "decrease".*

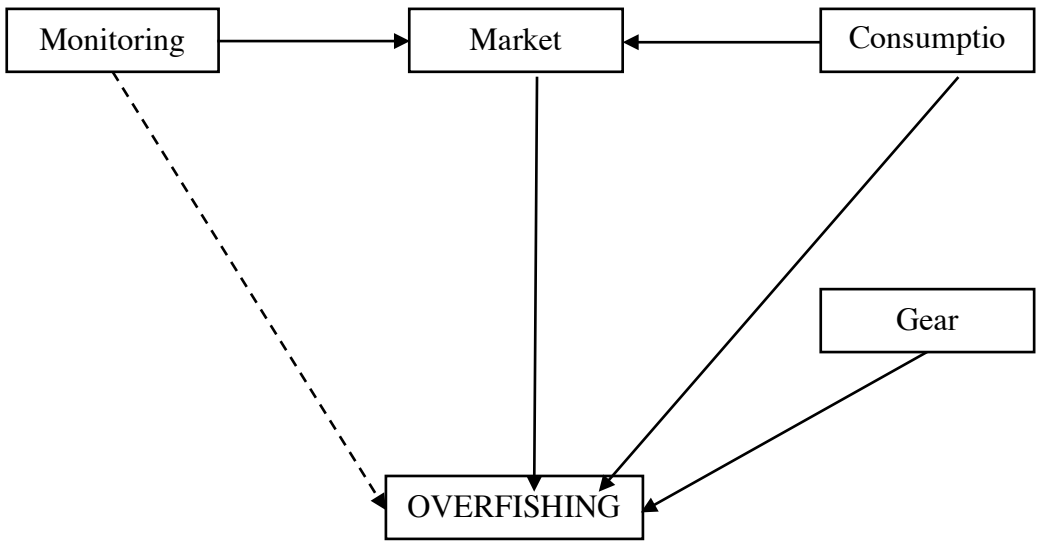
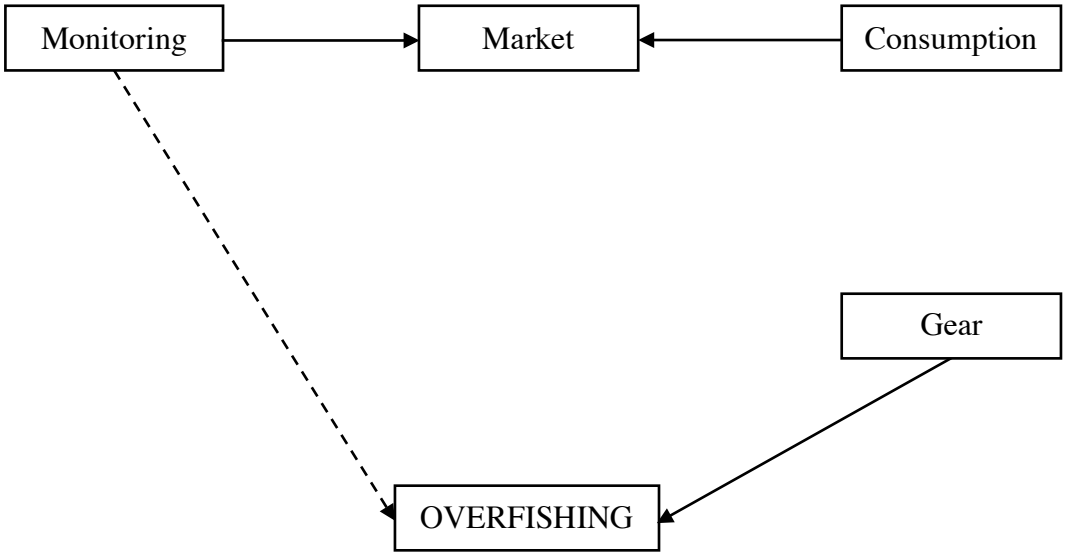


Figure 5. *Consensual Causal Network created using IEA, Experiment 2. Dashed line indicates a causal relation labeled as "decrease".*



The number of participants representing each path in the resulting consensual networks, as well as the judged mean strength, are reported in Table 7

Table 7.

Paths and mean path strengths of consensual network created using MSC, Experiment 2.

Cause factor	Effect factor	N (%) representing the path	Mean path strength (SD)
Monitoring	Overfishing	24 (80%)	-0.66 (0.24)
Monitoring	Market	22 (73.3%)	0.6 (0.21)
Consumption	Market	21 (70%)	0.86 (0.22)
Gear	Overfishing	21 (70%)	0.69 (0.27)
Consumption	Overfishing	20 (66.7%)	0.74 (0.28)
Market	Overfishing	19 (63.3)	0.82 (26.4)

Causal models and counterfactual judgments

As in Experiment 1, the total path strength of each of the factors was calculated for each participant. These are reported in Table 8. These strengths were then correlated with the corresponding counterfactual judgments. Table 9 displays the mean (SD) counterfactual judgments.

In Experiment 2, the mean correlation across individuals (i.e. the mean of a set of r correlations), was 0.66 (SD = 0.55, N = 33), accounting for 44% of the variance in the counterfactual judgments. This correlation was significantly different from zero, p<0.001. A paired samples *t* test was then used to calculate the difference between the mean total path correlation (r_m = .66, *SD* = .51) and the mean direct path correlation (r_m= .55, *SD* = .57). This difference was significant: *t* (32)= 2.89, *p* < .001, suggesting that the variance in individuals' counterfactual judgments is better explained by looking at the factors' total path strength (sum of direct and indirect paths), as opposed to looking at the direct path strengths alone. On the other hand, the difference between the mean total path correlation and the mean indirect path correlation (r_m=. 60, *SD* = .48) was not significant: *t*(32)= .68, *p* > .05. The correlation between judgments and number of paths emanating from each factor was -0.44 (SD=0.56).

In contrast to Experiment 1, an independent samples *t* test revealed that total path correlations were affected by whether participants completed the diagram task before or after the counterfactual judgment task. The group completing the diagram task first had significantly higher correlations (r_m= .89, *SD* = .19, *N* = 16) than the group completing the counterfactual judgment task first (r_m= .44, *SD* = .62, *N* = 17): *t*(31)=-2.82, *p* < .05. To investigate this further, a mixed model analysis of variance was carried out with the type of correlation as the within subjects dependent variable

and task order as the between subjects dependent variable. Both the type of correlation and the task order were significant, but there was no interaction between the two.

Table 8.

Mean (SD) total path strengths from each factor to overfishing, Experiment 2.

Factor	Mean (SD)
Consumption	0.9 (0.73)
Market	0.87 (0.5)
Monitoring	-1.36 (1.29)
Gear	0.58 (0.73)

Table 9.

Mean (SD) counterfactual judgments, Experiment 2. Judgments based on the decrease frame have been reversed to compute the means.

Factor	Mean (SD)
Consumption	31.21% (23.8)
Market	27.42% (18.81)
Monitoring	-35.06% (26.74)
Gear	12.93% (31.87)

On the other hand, counterfactual judgments were generally unaffected by task order with the exception of *gear*: the mean change rating of participants who completed the diagram first (M = 28.43, SD = 25.86) was significantly higher than that of participants who completed the change judgment task prior to the diagram task

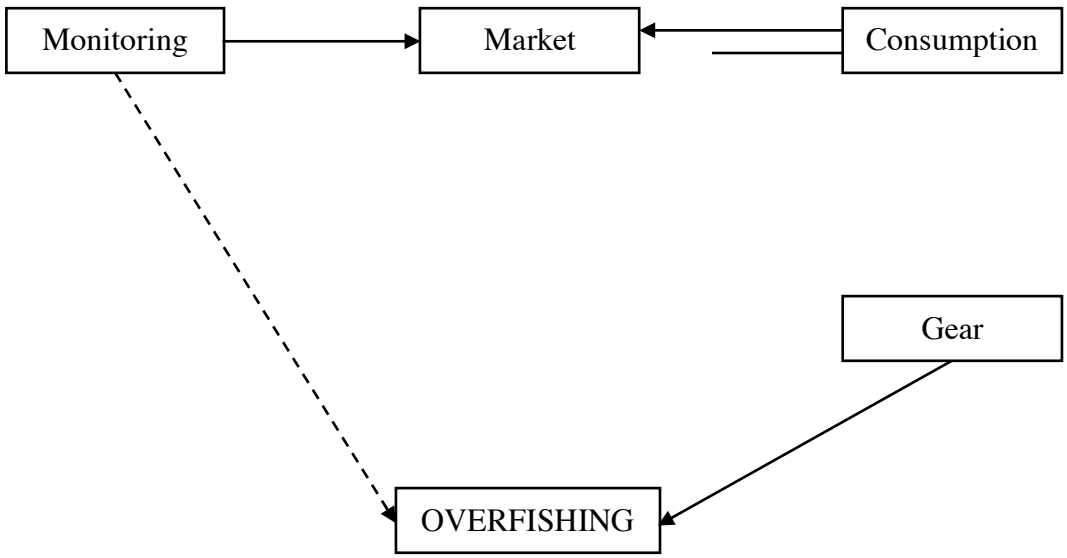
(M = -1.64, SD = 30.62): $t(31)$= -3.03, $p < .01$. There was no effect of task frame on total path correlations: $t(31)$= -1.636, $p > .05$, nor on any of the counterfactual judgments (all $ps > .05$). NEP scores did not vary as a function of individual causal models or counterfactual judgments.

Analysis of feedback loops

Each individual causal network was inspected for the presence of feedback loops. 52% of networks had at least one feedback loop. An example of an individual network is shown in Figure 6.

The mean number of loops across all networks was 1.58 (SD=2.3). The mean number of loops across only the networks containing loops was 3 (SD=2.48), ranging from 1 to 10 loops per network. Loops involved 2 to 5 factors. As in Experiment 1, the most common loop was *market-consumption* (N=9).

Figure 6. *Example of an individual network with a feedback loop, Experiment 2.*



The sample was then divided into two groups based on the presence of the *market-consumption* loop. An independent samples *t* test was used to compare reported fish consumption in the group that constructed the loop ($N = 9$, $M = 2.44$, $SD = 2.23$) and the group that did not represent the loop ($N = 24$, $M = 6.13$, $SD = 5.78$). A significant difference in fish consumption was found: $t(31) = -2.635$, $p = .01$ (equal variances not assumed), suggesting that participants who included the *market-consumption* loop consumed fish less often than the participants who had not represented it in their diagrams. Across the general sample, consumption varied from 0 to 20 times per month ($M = 5.12$, $SD = 5.29$).

CFC-14

The Immediate factor of the CFC-14 scale appeared to have good internal consistency, Cronbach's $\alpha = .85$. All items correlated well to the total factor (lowest $r = .50$). The Future factor had an acceptable internal consistency, $\alpha = .76$. The reliability of the scale could be increased if item 7 was removed. This item had a low correlation with the total factor ($r = .28$) but removing it would only increase Cronbach's α by .01. Thus, all items were kept. A median split was used to divide participants in high Future and low Future groups.

There was a positive correlation between total path strength correlations and CFC-Future scores: $r(31) = .41$, $p < .05$, suggesting that the higher an individual scored on the CFC-Future factor, the higher the correlation between their network and their change judgments. There was no relationship between an individual's CFC-Future score and the number of loops s/he constructed: $r(31) = .27$, $p > .05$.

3. 8 General Discussion

Summary of findings

The leading goal of the present research was to investigate how lay causal models of an environmental problem are related to reasoning about the issue. The first aim was to extend work by Green et al. (1995, 1998) by showing, in an environmental domain, that a person's causal network diagram correlates with their ratings of the effectiveness of actions based on these factors. In Experiment 1 and 2, participants completed two main tasks. The causal diagram task involved drawing a network diagram of how a set of factors related to overfishing may affect overfishing and each other. The counterfactual task consisted in judging the effectiveness of a series of counterfactual suppositions, based on the diagrammed factors, in reducing overfishing. Both experiments explored the relation between these two tasks. In line with our experimental hypotheses, total path strength of the diagrammed factors correlated more significantly with the counterfactual judgments than direct path strengths alone, or simply the number of paths emanating from a factor. In both experiments, total path strength was found to explain 44% of the variance in the counterfactual judgments. In addition, Experiment 2 found participants who had a high score on a scale measuring concern with future consequences, had significantly higher correlations between the diagram and the counterfactual judgments, than the group who had a low score.

The second aim was to investigate the extent to which people think about overfishing in a unidirectional way. Previous studies by White (e.g. 2008) analyzed participants' consensual network of forest ecosystems and found no feedback loops. In contrast, Experiment 1 and 2 analyzed individual networks and found that over half

of the participants built loops into their network diagrams. These loops varied in factors and sizes. In addition, Experiment 2 showed that participants who drew a feedback loop involving unsustainable fish consumption and presence of unsustainable fish on the market, reported consuming significantly less fish than the group who did not represent that specific loop.

Theoretical implications

The present findings have major theoretical implications for two domains. The first of these disciplines is that of cognitive science, aiming to elucidate the processes underlying general causal reasoning. The second field is that of environmental psychology, as well as the specific phenomenon of overfishing. Theoretical implications will be discussed in respect to each of these fields.

Causal reasoning. Experiment 1 and 2 showed that total path strength of the diagrammed factors correlated more significantly with the counterfactual judgments than direct path strengths alone, or simply the number of paths emanating from a factor. The first thing this implies is that when people engage in causal reasoning about a phenomenon, they can and do recruit a whole causal model as opposed to just individual direct causal relations (Lagnado et al., 2007). Naturally, the current experiments utilized only five and four factors, so it is difficult to generalize this implication to situations involving more variables. When reasoning about more factors, participants would increase the load on working memory – this might result in people resorting to a more simplified strategy based on explicit representation of direct path strengths only. Participants in the Green et al. (1995; 1998) studies represented up to twelve factors, therefore suggesting that at least with that many

factors, people can recruit a holistic causal representation of the phenomenon in question.

The current findings clearly establish that it is the strength of the causal paths as opposed to the sheer presence of them that is important. However, the present research cannot elucidate on the cognitive mechanisms that operate when people reason and combine the strengths of direct and indirect causal relations. Total path strengths were calculated by summing the strengths of direct and indirect paths, whilst indirect path strength was calculated by multiplying the strengths of indirect paths. This formula is intuitive because it weighs direct paths more than indirect paths. Similarly, it accounts for the fact that the weight of an indirect path decreases as the number of indirect paths increases. In other words, as the number of steps that it takes to get from a cause to an effect increases, the importance of each of these steps decreases. This method of computing total paths strengths explains only 44% of variance in the counterfactual judgments. A different algorithm for computing total path strengths might provide a better account. However, even though the present research adopted a quantitative approach in extracting and analyzing causal representations, people's spontaneous representation of causal relations might be qualitative (Lagnado, 2011; Pearl, 2000). In other words, even though it is clear that people take causal strengths into consideration, they do not need to have access to their precise values.

The second finding with implications for causal reasoning is that in both Experiment 1 and 2, participants constructed feedback loops in their causal diagrams. On the surface these findings appear to be in direct contrast with previous research by White (e.g. 2008). However, the two cannot be directly compared as the present study adopted a novel experimental approach that involved analyzing individual

causal networks as opposed to the consensual network (the current study also used different causal factors and a different causal network elicitation method). Inherently, this implies there are significant individual differences in causal reasoning that, more often than not, get neglected at an experimental level. Future studies should aim to integrate individual differences as part of their approach.

The presence of feedback loops suggests that people can appreciate two-way causal relations within a complex network. This notion is reinforced by the finding that, in both experiments, the most popular loop was also the most intuitively sensible one: most participants who constructed loops had a bidirectional link from consumption of unsustainable fish to the extent to which unsustainable fish is sold on the market. Furthermore, in Experiment 2, the group representing this loop reported consuming less fish than the group who did not represent that loop. This finding can be taken as preliminary evidence that people not only represent loops, but may also integrate them into their reasoning. Either way, further analyses would have to determine if a causal model involving loops provides a more accurate account of causal judgments than a model without loops. Similarly, further work is needed to explore the connection between the structure of individual causal representations and actual consumer behaviour.

The second aim of Experiment 2 was to investigate the nature of the individual differences in the cognitive mechanisms that might underlie construction of feedback loops. It explored the idea that participants who construct feedback loops might unfold causal chains over time and therefore they might have greater inclination to think about the future (as measured by the CFC scale). However, there were no systematic differences in the causal diagrams produced by participants with higher CFC scores. No differences were found in terms of number or types of loops, or total

number of links. Further investigation is needed to decipher what construct might lie at the core of individual differences in two-way causal reasoning. On the other hand, participants with higher CFC scores had significantly greater total path strength correlations between the causal diagram and the counterfactual judgments. This difference might be explained by the possibility that future-oriented individuals simply invested more cognitive effort across the two tasks (Nowack, Milfont and Van der Meer, 2012.)

Environmental psychology and overfishing. Inspection of the individual causal diagrams, as well as the consensual network, revealed some interesting patterns that have potentially important implications for overfishing. In both experiments, the majority of participants had a direct link from *consumption* to overfishing (and from *demand* to overfishing in Experiment 1).

Consumption is also understood to affect overfishing through *market* (majority had link from *consumption* to *market* and from *market* to overfishing). However, what is perhaps more striking is that only a minority recognizes that *consumption* and *market* may influence *monitoring* and *gear*. In other words, people seem to think that the extent to which the government monitors and enforces fishing laws, and the extent to which fishermen use destructive fishing gear, is not contingent on the consumption rate which sets market targets. This implies people are likely to view monitoring and gear as factors beyond their control. If people attribute responsibility to multiple factors they do not view as causally linked to them, this could result in the classic bystander effect (Darely & Latane, 1968) whereby intention to act decreases as shared responsibility for a problem increases. This is in line with Belk, Painter and Semenik's (1981) finding that participants who attributed an energy shortage to non-

personal causal factors (e.g. government, oil companies) also tended to favor a non-personal solution to the problem.

More generally, the current study has important implications for the domain of environmental psychology. First of all, it highlights the need to incorporate a causal approach in its endeavors, both as a theoretical contribution and as an experimental method. Numerous theoretical frameworks have been developed to explain the gap between the possession of environmental knowledge or awareness and displaying pro-environmental behavior. Although many hundreds of studies have been done, a comprehensive model is yet to be found (Kollmuss & Agyeman, 2002). Most existing theories adopted a social-psychological approach - examples are cognitive dissonance theory (Thogersen, 2004), norm-activation theory (Stern et al., 1999), and the theory of planned behavior (Ajzen, 1991). None of these frameworks are based on or include causal reasoning. Ajzen's theory of planned behaviour (TPB) seems to be the most widely used model to predict or explain variance in pro-environmental behaviour. Broadly speaking, TPB postulates that an intention to act environmentally is formed in a rational choice process weighting three different aspects: the person's attitudes towards the behaviour, the person's perception of social pressure to act in a certain way and the person' perception of behavioral control in the situation. Even though a person's perception of behavioral control is bound to be related to their causal model of the problem, the theory does not make explicit reference to causal models. As shown by the current findings, there are significant individual differences in causal models of environmental problems. Therefore, TPB is likely to provide a better account of pro-environmental behaviour by factoring in an extension that considers causal understanding of the problem.

In terms of experimental methodologies, past research in environmental psychology has looked mainly at attitudes towards nature (e.g. Milfont & Duckitt, 2010), pro-environmental intentions (e.g. Bamberg & Möser, 2007) and a relatively limited range of pro-environmental behaviors such as recycling and energy saving. With few exceptions, previous work has relied on self-report tools to measure attitudes, intentions and behaviors. Environmental knowledge has also been assessed through using multiple-choice questionnaires, surveys and factual tests. The present study points out how much can be learned by relying on a causal network method that retrieves environmental beliefs as a function of an interconnected network.

Limitations

There are some shortcomings with the current network elicitation method. The main problem concerns the method of selection of the factors to include. The factors in the current study were selected based on expert assessment of the relative importance of different factors. The problem this can pose for a causal network study is that it omits causally relevant factors that may either alleviate overfishing or act against the factors that are judged to cause overfishing. The result, might be a network representing a kind of vicious circle in which negative factors all interact to aggravate each other (White, 1995). The second problem for this method is that the factors selected do not constitute a closed system. In a way, this problem is unavoidable because there is no such thing as causally closed system. Therefore, the structural characteristics of the network may be affected by the omission of causally relevant items. Finally, a potential problem of employing a diagram to extract causal network is that participants may avoid creating too many crisscrossing patterns or drawing too many arrows for representing links and loops. Therefore, networks may

be left simplified and or incomplete in the pursuit of avoiding visual clutter, rather than due to lack of awareness of causal paths.

Practical implications

Understanding the nature of public beliefs of the risk factors and prevention of overfishing and other environmental problems is critical to the design of communication programs aimed at altering actions. In particular, such programs need to ensure consumers understand how their purchasing choices are causally linked to overfishing. However, as indicated by the current findings, it is not sufficient to make people aware of the presence of a causal link between them and the undesired effect. Evidently, it is imperative to emphasize the causal strength of that link. Most campaigns simply tell consumers that, along with seafood retailers and restaurants, they play a crucial role in the conservation of ocean resources. However, they do not quantify that role, they do not make it explicit that if consumers made the right choices, overfishing would not take place. As argued previously, the quantification does not necessarily need to come in a numerical format, but should convey an idea of magnitude of relation of some sort. The current findings also provide evidence that people can understand feedback loops, implying that environmental campaigns should also emphasize the cyclical nature of the natural ecosystem. Attention should be drawn on repercussions of unsustainable actions.

Future research

Future research should aim to extend the current findings to causal models with more factors and ideally, idiosyncratic factors. Comparison of causal models of different groups, such as experts versus non experts, consumers versus fishermen,

might also help shed light on the nature of lay understanding of environmental problems and related causal reasoning.

From a more cognitive perspective, it would be important to explore how people understand the different ways in which causes combine to bring about an effect. There are several different functions that can mediate the impact of each individual cause into the final effect (Steiner, 1972). One of the difficulties in allocating causality arises from the fact that causes can combine in various different ways to bring about an outcome. Three common functions are addition, conjunction or disjunction. In the additive case, each cause contributes something to the final outcome. For example, using destructive fishing gear and surpassing fishing quotas both contribute to overfishing. In the conjunctive case, all causes need to surpass a certain threshold. For example, fishing gear can only affect overfishing if the government does not monitor and enforce fishing laws properly. In the disjunctive case, it only takes one cause to bring about the outcome. People's consumption of unsustainable fish is a prime example – if nobody consumed unsustainable fish, overfishing would not take place. Future research should explore how sensitive people are to this reasoning and especially how it mediates attributions of responsibility (Gerstenberg & Lagnado, 2010).

Conclusion

To conclude, this research shows how naïve causal models of a complex environmental problem might hold the key to unlocking the reasoning processes underlying people's decisions to support sustainable behaviour. Environmental psychology and ecological campaigns can achieve greater success in their pursuits by making causal reasoning one of their main driving forces.

Chapter 4: Reasoning with causal loops: understanding everything.

4.1 Introduction

Background

The Earth's environment is a network of complex dynamic systems composed of interdependent chains of cause and effect. These dynamic systems include environmental systems such as the Earth's climate and food webs, financial systems such as the world's economies and businesses and social systems such as a country's government and culture. The interdependence characterizing all these systems can be grasped in terms of causal loops – feedback with and between systems. In other words, a system's output serves as input to that same system, or another system, therefore creating a circular process.

Some of these loops stabilize the system ('negative' causal loops), such as predator-prey relationships: an increase in the number of preys leads to an increase in the number of predators, which will lead to a decrease in the number of preys. Other examples of negative loops include the water cycle, the carbon cycle and human body homeostasis. Other loops drive the system in one direction ('positive' causal loops), such as global warming: an increase in the atmosphere's temperature leads to an increase in evaporation, which will lead to an increase in global warming. Other examples of positive loops are human population growth, economic recession and depletion of fisheries.

It is sufficient to consider some of the examples of loops listed above to appreciate the interdependency between the Earth's systems and how most global

problems are interconnected and exacerbate each other. An increase in human population leads to an increase in global emissions as well as putting further pressure on the world's resources such as fisheries. Global warming and resource depletion create and exacerbate a financial crisis. An increase in global warming leads to an increase in resource depletion, which in turn leads to further global warming. The circular structure of these system means the connections are potentially endless.

This interlinking of issues, or complex interdependency of problems, has implications both for the way people think about these issues – their forms of knowledge – and for the way they might go about beginning to solve them. The question then becomes if all these problems are ultimately just different facets of one single crisis, a crisis of causal reasoning. Causal reasoning is the ability to identify relationships between events or forces in the environment (causes) and the effects they produce. Therefore people use causal cues and their related effects to make decisions efficiently, to make predictions about the future circumstances of their environment and to fully understand mechanisms leading to change. Consequently, the key question is can (and do) people make appropriate causal inferences based on the cyclical convoluted structure of the world and its issues? The current study aims to explore the answer to this question.

Representation of causal loops

Before investigating people's ability to reason with causal loops, it is important to establish whether people actually construct and represent such loops in their own causal models of the world. Past research in causal cognition has already explored complex causal structures (Lagnado, Waldmann, Hagmayer, & Sloman, 2007; Waldmann, Hagmayer & Blaisdell, 2006; for a review see Rottman & Hastie,

2013). A central finding from this research is that people do indeed create and use causal models to structure their learning and inference. However, only a few studies have looked at the role of causal loops (e.g. Kim, Luhmann & Ryan, 2009) or dynamical systems (e.g. Rottman & Keil, 2011).

Given the ubiquity of causal cycles in the world, from an intuitive perspective, it would seem plausible to assume people reason about causally related events that occur in cycles (Rehder & Martin, 2011). For example, in economics, people expect that an increase in corporate hiring may increase consumers' income and thus their demand for products, leading to a further increase in hiring. Similarly, in meteorology, people expect that melting tundra due to global warming may release the greenhouse gas methane, leading to yet further warming. Following this idea, a series of studies investigating people's own real-life causal theories (e.g. Hagmayer & de Kwaadsteniet, 2008; Kim & Ahn, 2002a, 2002b; Rein, Love, & Markman, 2007; Sloman et al., 1998) have found that, in addition to linear causal structures, people also commonly report causal cycles. Kim (2005) asked lay people to consider mental disorders and found that 65% of participants spontaneously reported causal cycles.

Relatedly, in a study exploring people's ability to think in terms of two-way causal relations, Green (2001) investigated people's ability to create interactive models of ecological systems. He presented participants with a food web and asked participants to explain a complex pattern of fluctuation over time in the population of an herbivore. He found that most people were able to construct interactive accounts involving two, and in some cases three, entities (plant, herbivore and carnivore). However, as argued by White (2008), the system, or food web, comprised only three entities and individuals were constrained to explain a complex pattern presented to them, rather than envisaging themselves what sort of pattern might occur.

Following this idea, Nikolic and Lagnado (under review) investigated how lay representations of the causes of a complex environmental problem may underlie individuals' reasoning about the issue. They derived a set of factors in relation to overfishing (e.g. consumption of unsustainable fish, market selling unsustainable fish, destructive fishing practices). In two experiments participants were asked to draw diagrams showing how these factors, or causes, were linked to overfishing and to each other. Analyses based on individual causal networks diagrams revealed the presence of numerous feedback loops. Nikolic and Lagnado found that 52% of participants drew at least one causal loop in their diagram (with a an average of 3 loops per diagram). Furthermore, these loops often comprised over two factors. For example, people often connected an increase in consumption of unsustainable fish leading to an increase in the market selling unsustainable fish, which in turn leads to an increase in consumption.

As to how exactly people represent causal loops, Kim, Luhman and Ryan (2009) argued that people reason with a simplified representation of causal loops. They posit people represent causal loops as causal chains extending one step into the future. This simplified representation captures the loop as simply as possible while still maintaining the essential nature of the loop (Rehder & Martin, 2011). They provide two reasons for this assumption. First, because variables rarely cause each other constantly and simultaneously, it is likely that people assume that they influence each other in discrete time steps. They make their point using an example of a causal loop involving insomnia and poor school performance. It seems unlikely that people would think that a student's school performance is actually deteriorating as he or she sits up at night. Instead, it seems more realistic that the student's insomnia, say on Monday night leads to poor performance on Tuesday, which would then lead to a

sleepless Tuesday night, and so on. This example demonstrates how causal loops might be more realistically represented as causal chains that play out over time. Such a chain would allow Factor A to influence Factor B at Time 1, Factor B to influence Factor A at Time 2, and so on. Second, because it is implausible that people represent time steps extending into infinity, only a limited number of steps are likely to be considered. Kim et al. found support for their hypothesis through a series of experiments investigating the importance (or 'conceptual centrality'; Sloman, Love, & Ahn, 1998) people assign to factors involved in causal loops as opposed to linear causal structures (e.g. causal chains). They found that participants' judgments indicated that they did not consider loop factors to be the most central to the underlying concept. In other words, people seemed to unpack causal loops into causal chains.

Reasoning with causal loops

Taken together, these findings clearly show that laypeople spontaneously construct causal loops into their causal models. However, what these findings cannot elucidate is to what extent people make appropriate causal inferences based on these loops. Representation of a causal structure does not necessarily imply ability to reason about it.

Conversely, it is precisely this ability to reason with such causal structures that may be part of the key to understanding and solving many of the world's problems. This is because, as exemplified earlier, the world's problems are defined and fueled by vicious cycles. In democracies, the beliefs of the public, not only those of experts, affect government policy. Therefore effective risk communication is grounded in deep understanding of the mental models of policy-makers and citizens. Of particular

relevance in this day and age, is the fact that if people believe that anything can be done to nature, without repercussions for the human world, they are less likely to exhibit sustainable behavior. This follows from the idea that how humans treat the world must to some extent reflect what they believe about the effects of that treatment. Kempton (1986) pointed out that lay models about physical systems influence real life decision-making. He found that people's mental models of thermostats accounted for how they treat the control of heat in their homes. Those who possessed one theory tended to behave more economically than those who possessed the other theory. Kempton (1986) proposed, on the basis of interviews, that people used two distinct models of home heating systems. In the (incorrect) valve model, the thermostat is thought to regulate the rate at which the furnace produces heat. Therefore setting higher makes the furnace work harder. In the (correct) threshold model, the thermostat is viewed as setting the goal temperature, but not as controlling the rate of heating. Hence the furnace runs at a constant rate. Kempton then examined thermostat records from real households and found that the patterns of thermostat settings fitted nicely with the two models he had found. As another example, Atran, Medin, and Ross (2005) found that cultural groups' mental models of plant/animal interactions in the rainforest were consistent with the environmental impact of those groups. Therefore, common-sense understanding of the structure of cause and effect in nature, with specific focus on understanding of cyclical interactions between humans and nature, is an important topic from both a scientific and conservationist point of view.

Given the significance and urgency of the matter it is surprising that to date there have not been investigations focused specifically on exploring how people make causal inferences based on causal loops. However, the system dynamics domain has a

lot to say about how people's decisions might be influenced by cyclical structures. Sterman (e.g. Sterman and Booth Sweeney, 2002), as well as numerous others (e.g. Moxnes, 2000), used typical system dynamic tasks to investigate people's ability, or better inability, to make decisions based on complex dynamic systems - settings with multiple feedback loops, time delays, and nonlinearities. These studies seem to suggest that even well educated individuals generally lack the cognitive skills necessary for understanding the behavior of these systems Specifically, these studies propose that people display a general tendency to think in terms of linear causal structures and therefore fail to perceive key causal loops.

In the classic study, Sterman (1989b) examined a simple inventory management task, the "beer distribution game," in which participants were asked to minimize costs as they managed the production and distribution of a commodity. Though simplified compared to real firms, the task was dynamically complex as it included multiple feedbacks, time delays, nonlinearities, and accumulations. He found that participants generated costly oscillations with consistent amplitude and phase relations, even though the demand for the product was essentially constant. Importantly, econometric analysis of participants' decisions showed that people were quite insensitive to the feedback loops and time delays in the system. Sterman (1989a) found subjects exhibited the same behaviour in a simulated macroeconomy with time delays and feedback loops.

Given the alarming implications that these results have for the management of renewable resources, a series of studies have investigated participants' ability to manage common resource pools such as fish stocks (Moxnes, 1998). Moxnes found that subjects consistently overinvested leading to overexploitation of the resource in question. People's behaviour has been explained (Sterman, 1989a, 1989b) by

"misperception of feedback", or, in other words, linear causal thinking. The basic idea is that people generally adopt an event-based "open loop" view of causality, therefore ignoring feedback processes and time delays. Inevitably, this generates systematic dysfunctional behavior in the presence of dynamic complexity. Similar conclusions have been drawn from studies in experimental economics, psychology and management (Smith, Suchanek & Williams, 1988; Funke, 1991; Brehmer, 1992).

Current study

Given the aforementioned studies and discussion we can draw three implications. The first one is that it is clear that people do construct causal loops to make sense of the world (e.g. Nikolic & Lagnado, under review). The second claim is that even if people do represent causal loops in their own models, when they are asked to make decisions based on the world's cyclical systems, their reasoning starts to break down (e.g. Sterman). Allegedly, this happens precisely because of failure to appreciate causal loops. Thirdly and perhaps most undeniably, if people cannot reason properly with causal loops and make decisions based on the systems that nest these loops, then the world might be in trouble. In other words, solving current problems such climate change may be more challenging and new problems are bound to arise from the existing ones.

Given these three implications, the aim of the current study is to investigate the inevitable question of when is it exactly that human reasoning begins to crumble in the face of a causal loop. In order to really start addressing this question it is important to take a step back to basics. Therefore, the current study will adopt a bottom-up approach to investigate if people can engage in basic forms of reasoning - proper causal inferences - based on simple representations of causal loops.

Before proceeding to describe the current study it is important to discuss what constitutes a 'proper' causal inference. What distinguishes a true causal inference from a mere estimate of covariation is that the former is sensitive to the difference between predictions based on merely observed events and predictions based on the very same states of events generated by means of intervention (Pearl, 2000, Spirites & Scheines, 2001). For example, observing the state of a barometer allows people to make predictions about the upcoming weather (observational inference), whereas manipulating the barometer does not license such a prediction (interventional inference). Whereas observational inferences allow people to capitalize on both causal and non-causal correlations, interventional predictions are based only on predictive causal relations.

This means that in everyday contexts, causal inferences are aided by manipulation of potential causes, or better, by people intervening on the world rather than just observing it. Mere observation can only reveal a correlation, not a causal relation. Thus causal induction in experimental science requires manipulation, control over an independent variable such that changes in its value will determine the value of the dependent variable whilst holding other relevant conditions constant. Everyday causal induction has these same requirements. Naturally, in a lot of cases, people already have some causal knowledge, so they can answer certain causal questions without actual intervention. On the other hand, there are many instances when causal inference can be difficult because it depends not only on what happened, but also on what might have happened (Lewis, 1986; Pearl, 2000; Sloman, 2005). In such cases, some of those questions can be answered through mental intervention, by imagining a counterfactual situation in which a variable is manipulated and determining the effects of change.

Accordingly, counterfactuals based on interventions and counterfactuals based on observations will also yield different causal inferences. People's sensitivity to this difference is a key concept framing the current study's experimental design. For example, the diagram in Figure 1 represents the causal relation between three factors: having a chest infection, having a cough and having a bad sleep (see Figure 1A). On this simple model, having a chest infection causes a cough that causes a bad sleep. Observing that a cough is present warrants the backwards inference that a chest infection is likely (see Figure 1B). In addition, it also warrants the forward inference that a bad sleep is likely. In terms of counterfactual observations, had one observed no cough, then one could have still made both a backwards inference and a forward inference. In other words, had there been no cough observed, then a chest infection would be unlikely and bad sleep would be unlikely. However, when reasoning about the consequences of a counterfactual intervention on an event, one should not change one's beliefs about the states of the normal causes of the event. Had one intervened on the cough by taking a cough medicine, the cough would not have been present (see Figure 1C). However, this would not warrant the backward inference that a chest infection would not have been present. This is a rational principle of inference because an effect is indeed not diagnostic of its causes whenever the effect is not being generated by those causes but instead by mental or physical intervention from outside the normal causal system. On the other hand, future inference is still possible: given the absence of cough, a bad sleep is unlikely.

Sloman and Lagnado (2005) investigated precisely how people's causal inferences might differ according to whether they are reasoning about a counterfactual observation or a counterfactual intervention. Their experiments confirm that causal inferences given counterfactual interventions are different from those based on

observations, because there is no backtracking with the former.

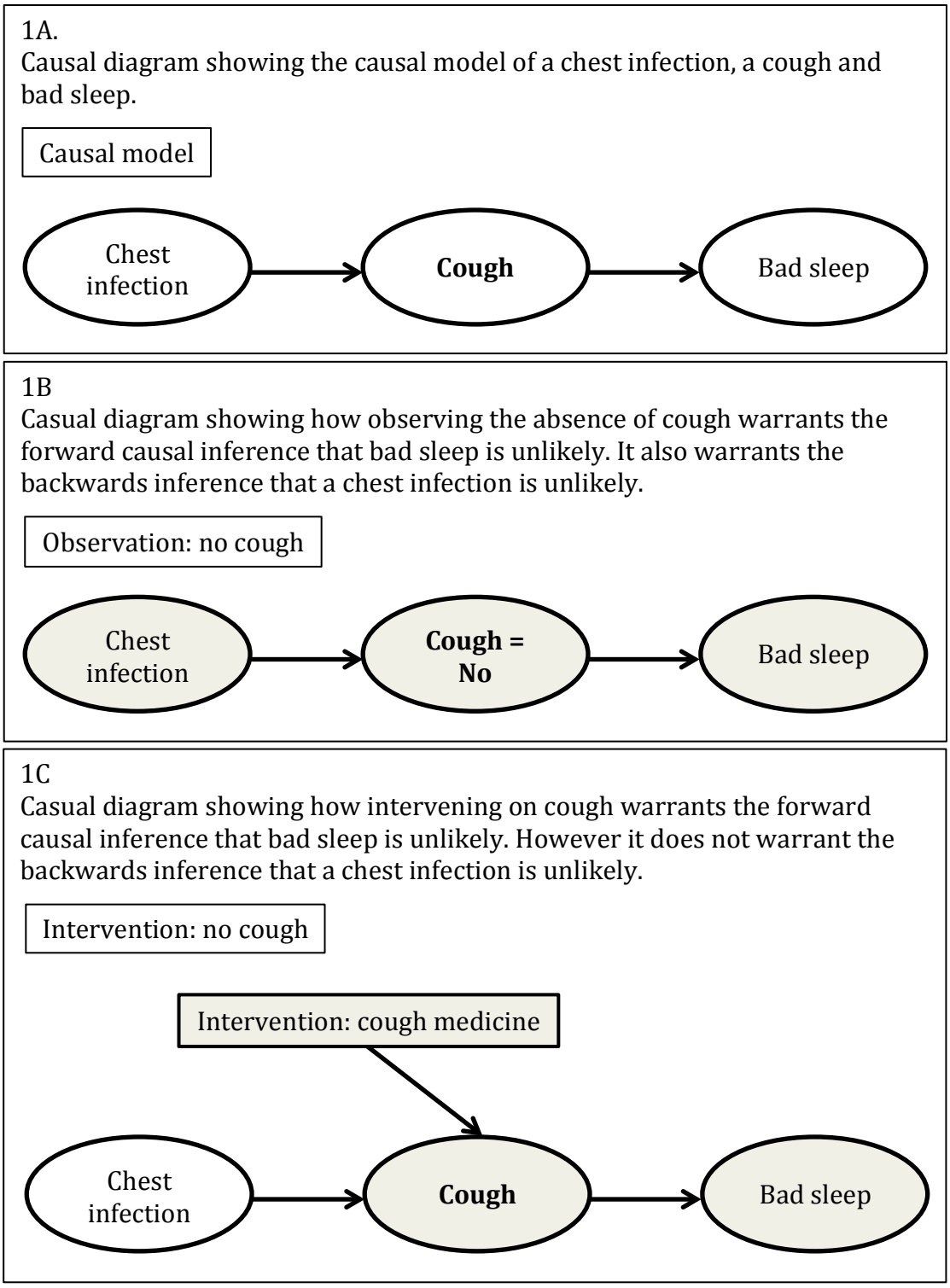


Figure 1. *Causal inferences warranted according to observations and interventions.*

4.2 Experiment 1

Participants were presented with a simple scenario based on a causal loop. The nature of the loop was counterbalanced so it was either a negative (stabilizing) or a positive (reinforcing) causal loop. Both types of loops were investigated because people might represent them differently, therefore yielding different causal judgments. For example, when it comes to positive loops, people might not represent them farther than a single time step into the future. This is because each factor becomes just a more extreme version of its previous past state.

The scenario was chosen to be a simple real life situation about sleeping patterns. This is because it was considered that people might have an easier time relating to a common pattern they are likely to have experienced first hand. The scenario based on the negative loop was about a person whose amount of sleep affected his day time level of tiredness, which in turn affected his amount of sleep (the more sleep the less tired, the less tired the less sleep, the less sleep the more tired, the more tired the more sleep and so on, therefore maintain an equilibrium). The scenario based on the positive loop was about a man who's amount of sleep affected his daytime levels of stress, which in turn affected his amount of sleep (the less sleep, the more stress, the more stress, the less sleep and so on, therefore creating a vicious cycle). Based on the idea that people represent causal loops by unpacking them into different time periods, participants were then provided with information about the state of the causal factors at five different time periods (in this case number of hours of sleep over five consecutive nights). The information matched the pattern described in the scenario.

Following the presentation of these values, participants were presented with either a counterfactual observation or a counterfactual intervention affecting the mid

time period (Night 3). In both conditions they were asked to imagine the person had slept a different number of hours than the one reported in the table. In the observation condition, participants were not given any reason why, other than asked to imagine the counterfactual supposition. In the intervention condition, however, they were told to imagine the person took a sleeping pill that affected their amount of sleep. At this stage it is important to acknowledge that the observation condition is not, strictly speaking, purely observational. This is because, one might argue, there is the worry that participants might assume that the change portrayed by the observation is in fact the result of some intervention. In other words, it is indeed possible that participants might perceive the counterfactual observation as slightly ambiguous. This point will be developed further in the Discussion of Experiment 2, as well as in the General Discussion.

In both conditions they were then asked to make causal inferences about the two past time periods (Night 1 and Night 2) and about the two future time periods (Night 4 and Night 5). The causal inference consisted in estimating how the number of hours of sleep may or may not have changed according to the counterfactual supposition. Therefore participants could choose one of the following answers: 'same', 'more' or 'less'. The answers required a qualitative causal inference (as opposed to an exact numerical estimate) in accordance with the idea that people's spontaneous representation of causal relations might be qualitative (Lagnado, 2011; Pearl, 2000).

The reason why participants were questioned about Time period 1 and Time period 5, (as well as Time period 2 and Time period 4), is to explore how people might or might not extend the backward and forward causal inferences to time periods further away from the time period affected by the counterfactual (Time period 3).

White, over a series of studies (1997, 1998, 1999, 2000), found that when participants are asked to judge the effects of a perturbation to one entity in a food web, people consistently judge that the greatest effects will be found for species immediately adjacent to the perturbed entity in the structure of the food web. Accordingly, he found that the magnitude of the effect rapidly drops off with increasing distance from the perturbation. In other words, what seems to be happening is that the causal effect is judged to be diminishing as a function of causal links between the point of 'change' and the target effect.

Green (2001) argued that this happens because individuals minimize what they represent explicitly - a basic supposition of the theory of mental models (Johnson-Laird & Byrne, 1991). This would mean that not all possible causal relations amongst the factors would be represented explicitly in their initial causal model. In this sense the theory is consistent with Kim and Ahn's (2009) hypothesis of how people represent loops – by unpacking them into basic causal chains. Therefore, it is possible that in the current experiment not all participants will extend their causal inferences further than one time point in the past and in the future. If so, this will provide support for Kim and Ahn's theory of representation of loops but also the idea that people reason counterfactually by running a mental model – that is mentally simulating the change.

The experimental hypothesis is that people are able to reason with simple causal loops and make sensible causal inferences accordingly. The extent to which people make sensible causal inferences when reasoning with causal loops can be formalized into three levels.

Level 1. The first level, and the stepping stone for more complex forms of reasoning, is simply differentiation between a true causal inference from a mere

estimate of covariation. As discussed above, this consists in sensitivity to the difference between predictions based on merely observed events and predictions based on the very same states of events generated by means of intervention. Accordingly, the first two experimental hypotheses will explore the extent to which people display level 1 causal reasoning.

Hypothesis 1: Participants will backtrack (change causal inferences according to the counterfactual supposition for both Night 2 and Night 1) only in the observation condition and not in the intervention condition.

Hypothesis 2: Participants will make the same forward inferences in both conditions (change causal inferences according to the counterfactual supposition for both Night 4 and Night 5).

Level 2. The second level consists in providing the correct normative inference. It is possible for participants to make different inferences given counterfactual interventions and observations, but they may not actually make the *correct* inferences. For example, they might backtrack only in the Observation condition, therefore grasping the presence of a truly causal relation, but perhaps not in the right direction. Similarly they could make similar forward inferences in both conditions, but these could be similarly wrong. This level can be conceptualized in the hypothesis below.

Hypothesis 3: The answers selected by participants will reflect the correct qualitative causal inference according to the counterfactual supposition.

Another component of making correct inferences is the degree to which these are extended to time periods further away from the perturbation (counterfactual supposition); i.e. Time period 1 and 5. Given the discussion above, the fourth

hypothesis follows.

> Hypothesis 4: Fewer participants will extend their causal inferences to Night 1 and Night 5.

Note hypothesis 4 is not a requirement for attaining level 2 causal reasoning, but a mere prediction of the kind of form reasoning might take given theories on representation of loops (e.g. Kim & Ahn, 2009).

Level 3. The final level of causal reasoning can perhaps be formalized as consistency within the model. In other words, the question is to what extent are people's own inferences coherent. Examining only the number of correct answers could be deceiving because the majority of people could indeed provide the correct answer for each time period, but this majority might not be the same group of people. In other words, participants can be deemed to have a full causal model of the loop only if their inferences are coherent with their own understanding of the causal loop. Hence, causal inferences about Night 1 should be contingent on causal inferences about Night 2; and causal inferences about Night 5 should be contingent on causal inferences about Night 4. Therefore, the fifth hypothesis follows below.

> Hypothesis 5: Participants' causal inferences about Night 1 and Night 5 will be conditional and consistent with their inferences about Night 2 and Night 4 respectively.

4. 3 Method

Participants

143 participants were recruited through Amazon Turk. Participation in the survey was limited to people living in the United States to maximize likelihood of

recruiting participants who speak English as their first language. The study was advertised as investigating reasoning about causes and effects. All participants were paid $0.50. 68 were males (48%) and 75 were females (52%). The mean age was 35.7 years (SD = 13.2; range 18 to73 years). The participants' education background was approximately equally split between Sciences (N=50; 35%), Arts (N=47; 33%) and Mixed (N=46; 32.2%).

Design

The experiment comprised two between-subjects conditions: the observation condition and the intervention condition. All participants were presented with both the positive loop scenario and the negative loops scenario (within-subjects). The order in which they completed the two scenarios was counterbalanced. The dependent variable was the answers to the four questions - a question about each time period. Participants were always questioned about the two past time periods first and the future ones after, but the order in which they were questioned about them was counterbalanced. The survey ended with a series of demographic questions.

Materials

The materials consisted of a web-based questionnaire. The first page of the questionnaire provided simple instructions telling participants they will be presented with two scenarios and asked to answer questions about them. Then, depending on the counterbalancing condition, participants were either presented either with the positive loop scenario followed by the negative loop scenario, or vice-versa. Following each scenario, participants were presented with the counterfactual manipulation (either based on an observation or an intervention according to condition). After the

manipulation participants were presented with the four causal inference questions. Finally, participants were given a series of demographic questions.

Scenarios. The two scenarios were both based on sleeping patterns and were designed to be similar to each other in terms of cover story. However, they were not designed to be directly comparable to each other.

Negative loop scenario

> When John sleeps 4 hours, he feels very tired the following day. When John feels very tired during the day, he sleeps 8 hours the following night. When John sleeps 8 hours, he feels very rested the following day. When John feels very rested during the day, he sleeps 4 hours the following night. The table below shows John's sleeping pattern across five nights.

Day / Night	**John**
Day 1	Very rested
Night 1	4 hours sleep
Day 2	Very tired
Night 2	8 hours sleep
Day 3	Very rested
Night 3	4 hours sleep
Day 4	Very tired
Night 4	8 hours sleep
Day 5	Very rested
Night 5	4 hours sleep

Positive loop scenario

When Pete sleeps 6 hours or less, he feels very stressed the following day. When Pete feels very stressed during the day, he has trouble sleeping the following night and sleeps one hour less than the previous night. When Pete sleeps 7 hours or more, he feels very relaxed the following day. When Pete feels very relaxed during the day, he sleeps 7 hours or more the following night.

Day / Night	**Pete**
Day 1	Very stressed
Night 1	6 hours sleep
Day 2	Very stressed
Night 2	5 hours sleep
Day 3	Very stressed
Night 3	4 hours sleep
Day 4	Very stressed
Night 4	3 hours sleep
Day 5	Very stressed
Night 5	2 hours sleep

Counterfactual manipulations. The counterfactual suppositions were identical for both scenarios (except the name of the character). The exact wording of the suppositions in the two conditions is reported below.

Intervention condition: Suppose that during night 3, contrary to what is stated above, Pete/John took a sleeping pill that made him sleep 8 hours (during

night 3), instead of 4 hours. The sleeping pill has no side effects other than increasing the amount of sleep for one night (night 3).

Observation condition: Suppose that during night 3, contrary to what is stated above, Pete/John actually slept 8 hours and not 4 hours.

Questions. The four causal inference questions were presented on two separate pages. The first page contained the two questions about the two time periods closest to the time period affected by the counterfactual supposition (Time 2 and Time 4). The second page contained the two questions about the two time periods further away from the time period affected by the counterfactual supposition (Time 1 and Time 5). The scenario and counterfactual supposition was displayed at the top of each page so participants could refer back to it. The questions all followed the same multiple-choice format. The exact wording for one of the questions is reported below.

E.g. Night 1: Given this new piece of information about night 3, what can you infer about the number of hours John slept during night 1?

- John would have slept less than 4 hours during night 1.
- John would have slept more than 4 hours during night 1.
- John would have slept the same number of hours during night 1.

The order in which the answer choices were presented was randomized.

Procedure

Surveys were programmed and administered through the Qualtrics web-based survey platform, hosted through an account licensed to University College London. At the end of the study participants were given the option to write any comments they may have had. There were no comments that indicated misunderstanding of the study.

The whole study lasted on average 10 minutes.

4.4 Results

The results for each scenario will be discussed separately. There was no effect of order of scenario presentation.

Negative loop scenario

Table 1 shows the percentage of participants (and number) in each answer category for the four questions, as a function of condition. The 'correct' normative answer for each question is indicated in bold.

Table 1

Percentage of participants (and number) in each answer category for the four questions, as a function of condition. Values highlighted in bold indicate the correct normative answers for each condition. Experiment 1, negative loop scenario.

Question	Answer	Intervention (N=74)	Observation (N=69)
Night 1	Less	4.1% (N=3)	5.8% (N=4)
	More	13.5% (N=10)	**62.3% (N=43)**
	Same	**82.4% (N=61)**	31.9% (N=22)
Night 2	Less	12.2% (N=9)	**73.9% (N=51)**
	More	6.8% (N=5)	4.3% (N=3)
	Same	**81.1% (N=60)**	21.7% (N=15)
Night 4	Less	**74.3% (N=55)**	**85.5% (N=59)**
	More	6.8% (N=5)	4.3% (N=3)
	Same	18.9% (N=14)	10.1% (N=7)
Night 5	Less	0% (N=0)	5.8% (N=4)
	More	**77% (N=57)**	**78.3% (N=54)**
	Same	23% (N=17)	15.9% (N=11)

The analyses will be discussed in respect to each of the five experimental hypotheses.

Hypothesis 1: Participants will backtrack (change causal inferences according to the counterfactual supposition for both Night 2 and Night 1) only in the Observation condition and not in the Intervention condition.

As can be observed in Table 1, most participants in the Intervention condition do not backtrack when asked about Night 2 (81.1%) or Night 1 (82.4%) - they state the number of hours of sleep would not have changed according to the counterfactual supposition. On the other hand, in the Observation condition, only a few answer 'same' when asked about Night 2 (21.7%) or Night 1 (31.9%). A chi-square for independence was run to compare the observed frequency of cases in each condition for both Night 2 and Night 1. As predicted by the hypothesis, there was a significant association between condition and answer choice for Night 2: χ^2 (2, n=143) = 56.8, p < 0.001, phi = 0.63; and for Night 1: χ^2 (2, n=143) = 38.9 p < 0.001, phi = 0.52.

Hypothesis 2: Participants will make the same forward inferences in both conditions (change causal inferences according to the counterfactual supposition for both Night 4 and Night 5).

As can be observed in Table 1, in the Intervention condition only a few participants state the number of hours of sleep would be the same at Night 4 (18.9%) and Night 5 (23%). A similar pattern of results emerges in the Observation condition for Night 4 (10.1%) and Night 5 (15.9%). A chi-square for independence was run to compare the observed frequency of cases in each condition for both Night 4 and Night 5. As predicted by the hypothesis, there was no significant association between condition and answer choice for Night 4: χ^2 (2, n=143) 2.80, p = 0.246, phi = 0.14;

nor for Night 5: χ^2 (2, n=143) = 5.2, p = 0.074, phi = 0.191.

Hypothesis 3: The answers selected by participants will reflect the correct qualitative causal inference according to the counterfactual supposition.

As can be observed in Table 1, the correct predicted normative answer is indicated in bold and the majority of participants select that answer for each time period and for both conditions. Given that there were three answer choices, the probability that the participant would choose the correct answer was 1/3. A binomial test to test was used to evaluate if the proportion of participants who selected the correct answer was significantly different from chance (for each time period and condition). In the Intervention condition, the difference was significant for all time periods. In all cases, the deviation from 0.3 was highly significant (binomial test, $p<0.001$). The same can be said about the Observation condition.

Hypothesis 4: Fewer participants will extend their causal inferences to Night 1 and Night 5.

Even though in the Observation condition the majority of participants selects the correct normative answer for both Night 2 and Night 1, fewer participants do so for Night 1 (62.3% select 'less) in comparison to Night 2 (73.9% select 'more'). Participants' answers were recoded into two categories: correct and incorrect. These were entered in a chi-square test for independence comparing Night 1 to Night 2. The chi-square revealed a significant difference between the two Nights: χ^2= (1, n=69) 4.81, p = 0.02.

The same can be said about Night 5. Even though in the Observation condition the majority of participants selects the correct normative answer for both Night 4 and

Night 5, fewer participants do so for Night 5 (78.3% select 'more') in comparison to Night 4 (85.5% select 'less'). However, this difference is not significant: $\chi^2(1, n=69) = 2.92$, $p = 0.09$.

When it comes to the Intervention condition, approximately the same number of people selects the correct answer for Night 2 (81.1% select 'same') and Night 1 (82.4% selects 'same'). The same can be said about the inferences about the future: a similar number selects the correct answer for Night 4 (74.3% select 'less') and for Night 5 (77% select 'more).

Hypothesis 5: Participants' causal inferences about Night 1 and Night 5 will be conditional and consistent with their inferences about Night 2 and Night 4 respectively.

Another analysis that will provide insight into participants' causal inferences is analyzing their answers for a specific time period as a function of their answers to other time periods – conditional analyses. This will reveal if people make correct causal inferences that are coherent with their own representation of the causal loop. Hence, causal inferences about Night 1 should be contingent on causal inferences about Night 2. Participants' answers for Night 1 given their answer for Night 2 are displayed in Table 2. The percentage indicates the proportion of participants who provided each conditional answer.

As can be seen in Table 2, for the Observation condition, the majority of participants who correctly infer that John will have slept fewer hours at Night 2 also infer that John would have slept more hours at Night 1 (N=56.5%). For the Intervention condition, the majority of participants who correctly infer that John will

have slept the same number of hours at Night 2 also infer that John would have slept the same number of hours at Night 1 (77%).

Table 2.

Table showing percentage of participants' answers to Night 1 as a function of their answer to Night 2. Values highlighted in bold indicate the correct normative answers for each condition. Experiment 1, negative loop scenario.

Answer for Time 2	Answer for Time 1	Intervention	Observation
Less	Less	1.4% (N=1)	4.3% (3=1)
	More	9.5% (N=7)	**56.5%** (N=39)
	Same	1.4%(N=1)	13.0%(N=9)
More	Less	2.7%(N=2)	0.0% (N=0)
	More	0.0% (N=0)	2.9%(N=2)
	Same	4.1% (N=3)	1.4% (N=1)
Same	Less	0.0% (N=0)	1.4% (N=1)
	More	4.1% (N=3)	2.9% (N=2)
	Same	**77.0%** (N=57)	17.4% (N=12)

Similarly, causal inferences about Night 5 should be contingent on causal inferences about Night 4. Participants' answers for Night 5 given their answer for Night 4 are displayed in Table 3. As can be seen in Table 3, for both conditions, the majority of participants who correctly infer that John will have slept fewer hours at Night 4 also infer that John would have slept more hours at Night 5 (Intervention condition: N = 66.2%; Observation condition: 71%).

Table 3.

Table showing percentage of participants' answers to Night 5 as a function of their answer to Night 4. Values highlighted in bold indicate the correct normative answers for each condition. Experiment 1, negative loop scenario.

Answer for Night 4	Answer for Night 5	Intervention	Observation
Less	Less	0.0% (N=0)	4.3% (N=3)
	More	**66.2% (N=49)**	**71.0% (N=49)**
	Same	8.1% (N=6)	10.1% (N=7)
More	Less	0.0% (N=0)	2.9% (N=2)
	More	4.1% (N=3)	1.4% (N=1)
	Same	2.7% (N=2)	0.0% (N=0)
Same	Less	0.0% (N=0)	0.0% (N=0)
	More	6.8% (N=5)	4.3% (N=3)
	Same	12.2% (N=9)	5.8% (N=0)

Positive loop scenario

Table 4 shows the percentage of participants (and number) in each answer category for the four questions, as a function of condition. The 'correct' normative answer for each question is indicated in bold.

Table 4.

Percentage of participants (and number) in each answer category for the four questions, as a function of condition. Values highlighted in bold indicate the correct normative answers for each condition. Experiment 1, positive loop scenario.

Question	Answer	Intervention (N=74)	Observation (N=69)
Night 1	Less	40.8% (N=8)	5.8% (N=4)
	More	9.5% (N=7)	**49.3% (N=34)**
	Same	**79.7% (N=59)**	41.3% (N=31)
Night 2	Less	14.9% (N=11)	5.8% (N=4)
	More	6.8% (N=5)	**59.4% (N=41)**
	Same	**78.4% (N=58)**	34.8% (N=24)
Night 4	Less	6.8% (N=5)	4.3% (N=3)
	More	**83.8% (N=62)**	**78.3% (N=54)**
	Same	9.5% (N=7)	17.4% (N=12)
Night 5	Less	4.1% (N=3)	4.3% (N=3)
	More	**82.4% (N=61)**	**85.5% (N=59)**
	Same	13.5% (N=10)	10.1% (N=7)

The analyses will be discussed in respect to each of the five experimental hypotheses.

Hypothesis 1: Participants will backtrack (change causal inferences according to the counterfactual supposition for both Night 2 and Night 1) only in the observation condition and not in the intervention condition.

As can be observed in Table 4, most participants in the Intervention condition do not backtrack when asked about Night 2 (78.4%) and Night 1 (79.7%) - they state the number of hours of sleep would not have changed according to the counterfactual

supposition. On the other hand, in the Observation condition, fewer participants answer 'same' when asked about Night 2 (34.8%) and Night 1 (41.3%). A chi-square for independence was run to compare the observed frequency of cases in each condition for both Night 2 and Night 1. As predicted by the hypothesis, there was a significant association between condition and answer choice for Night 2: χ^2 (2, n=143) = 45.4, $p < 0.001$, phi = 0.56; and for Night 1: χ^2 (2, n=143) = 27.7 $p < 0.001$, phi = 0.44.

Hypothesis 2: Participants will make the same forward inferences in both conditions (change causal inferences according to the counterfactual supposition for both Night 4 and Night 5).

As can be observed in Table 4, in the Intervention condition only a few participants state the number of hours of sleep would be the same at Night 4 (9.5%) and Night 5 (13.5%). A similar pattern of results can be observed in the Observation condition for Night 4 (17.4%) and Night 5 (10.1%). A chi-square for independence was run to compare the observed frequency of cases in each condition for both Night 4 and Night 5. As predicted by the hypothesis, there was no significant association between condition and answer choice for Night 4: χ^2 (2, n=143) 2.2, p = 0.334, phi = 0.12; nor for Night 5: χ^2 (2, n=143) = 0.388, p = 0.823, phi = 0.05.

Hypothesis 3: The answers selected by participants will reflect the correct qualitative causal inference according to the counterfactual supposition.

As can be observed in Table 4, the correct predicted normative answer is indicated in bold and the majority of participants select that answer for each time period and for both conditions. Given that there were three answer choices, the

probability that the participant would choose the correct answer was 1/3. A binomial test was used to evaluate if the proportion of participants who selected the correct answer was significantly different from chance (for each time period and condition).

In the Intervention condition, the difference was significant for all time periods (binomial test, $p<0.001$). However, when it comes to the Observation condition, there were a relatively large number of participants who did not backtrack (answered 'same') when asked about Night 1. This meant that even though the majority still picked the correct normative answer, this proportion was not significant for Night 1 ($p=0.231$). For all other time periods the difference is highly significant (binomial test, $p<0.001$).

Hypothesis 4: Fewer participants will extend their causal inferences to Night 1 and Night 5.

Even though in the Observation condition the majority of participants selects the correct normative answer for both Night 2 and Night 1, fewer participants do so for Night 1 (49.3% select 'more') in comparison to Night 2 (59.4% select 'more'). Participants' answers were recoded into two categories: correct and incorrect. These were entered in a chi-square test for independence comparing Night 1 to Night 2. However this difference is not significant: $\chi^2= (1, n=69)\ 2.95$, $p = 0.09$.

When it comes to forward inferences in the Observation condition, the majority of participants select the correct normative answer for both Night 4 and Night 5. However, in contrast to Night 1, slightly more participants do so for Night 5 (85.5% select 'more') in comparison to Night 4 (78.3% select 'more'). However, this difference is not significant: $\chi^2(1, n=69) = 2.13$, $p = 0.14$.

When it comes to the Intervention condition, approximately the same number of people selects the right answer for Night 2 (78.4% select 'same') and Night 1 (79.7% selects 'same'). The same can be said about forward inferences: a similar number selects the correct answer for Night 4 (83.8% select 'more') and for Night 5 (82.4% select 'more').

Hypothesis 5: Participants' causal inferences about Night 1 and Night 5 will be conditional and consistent with their inferences about Night 2 and Night 4 respectively.

Causal inferences about Night 1 should be contingent on causal inferences about Night 2. Participants' answers for Night 1 given their answer for Night 2 are displayed in Table 5. As can be seen in Table 5, for the Observation condition, the majority participants who correctly infer that Pete will have slept more hours at Night 2 also infer that Pete would have slept more hours at Night 1 (N=70.3%). For the Intervention condition, the majority of participants who correctly infer that John will have slept the same number of hours at Night 2 also infer that John would have slept the same number of hours at Night 1 (43.5%).

Similarly, causal inferences about Night 5 should be contingent on causal inferences about Night 4. Participants' answers for Night 5 given their answer for Night 4 are displayed in Table 6. As can be seen in Table 6, for both conditions, the majority of participants who correctly infer that Pete will have slept more hours at Night 4 also infer that Pete would have slept more hours at Night 5 (Intervention condition: N = 75.7%; Observation condition: 75.4%).

Table 5.

Table showing percentage of participants' answers to Night 1 as a function of their answer to Night 2. Values highlighted in bold indicate the correct normative answers. Experiment 1, positive loop scenario.

Answer for Night 2	Answer for Night 1	Intervention	Observation
Less	Less	5.4% (N=4)	2.9% (N=2)
	More	2.7% (N=2)	0.0% (N=0)
	Same	6.8% (N=5)	2.9% (N=2)
More	Less	0.0% (N=0)	0.0% (N=0)
	More	4.1% (N=3)	**43.5% (N=30)**
	Same	2.7% (N=2)	15.9% (N=11)
Same	Less	5.4% (N=4)	2.9% (N=2)
	More	2.7% (N=2)	5.8% (N=4)
	Same	**70.3% (N=52)**	26.1% (N=18)

Table 6.

Table showing percentage of participants' answers to Night 5 as a function of their answer to Night 4. Values highlighted in bold indicate the correct normative answers. Experiment 1, positive loop scenario.

Answer for Night 4	Answer for Night 5	Intervention	Observation
Less	Less	1.4% (N=1)	1.4% (N=1)
	More	4.1% (N=3)	2.9% (N=2)
	Same	1.4% (N=1)	0.0% (N=0)
More	Less	2.7% (N=12)	1.4% (N=1)
	More	**75.7% (N=56)**	**75.4% (N=52)**
	Same	5.4% (N=4)	1.4% (N=1)
Same	Less	0.0% (N=0)	1.4% (N=1)
	More	2.7% (N=2)	7.2% (N=5)
	Same	6.8% (N=5)	8.7% (N=6)

4.5 Discussion

Experiment 1 provided support for the hypothesis that people are able to reason with simple causal loops and make sensible causal inferences accordingly. Results are discussed briefly in relation to each loop scenario.

Negative loop

First and foremost, the results clearly show that participants were sensitive to the difference between predictions based on merely observed events and predictions based on the very same states of events generated by means of intervention. This means they displayed at least level 1 causal reasoning. Participants backtracked only in the Observation condition and not in the Intervention condition (hypothesis 1). The difference in backtracking between conditions was significant. Participants also made

similar forward inferences in both conditions– there was no significant differences between conditions (hypothesis 2). The second level of causal reasoning consisted in providing the correct normative inference. Indeed, the majority of participants made correct causal inferences for each time period and condition (hypothesis 3). The proportion of correct answers was significantly higher than chance for all time periods for both conditions.

In the Observation condition, participants gave significantly more correct responses for time period 2 rather than for time period 1. Participants also gave more correct responses for time period 4 rather than for time period 5, but this difference was not significant. On the other hand, in the Intervention condition, the proportion of correct responses was similar across all time periods. Finally, in order to display level 3 causal reasoning, participants had to make inferences that were coherent with their representation of the causal loop (hypothesis 5). This was indeed the case - conditional analyses showed that the majority of participants' causal inferences about Night 1 and Night 5 were conditional and consistent with their inferences about Night 2 and Night 4 respectively (this was true for both conditions).

Positive loop

A similar pattern of results emerged for the positive loop scenario: participants clearly displayed level 1 causal reasoning. In terms of level 2, the proportion of correct answers was significantly higher than chance for most time periods and for both conditions – however the difference was not significant for time period 1 in the Observation condition (there were a relatively large number of participants who did not backtrack). This apparent lack of backtracking in the positive loop scenario could be in line with the idea that people might not represent positive loops further than a

single time step into the past and into the future. However, participants did make correct forward inferences up to time period 5 so at this stage it is hard to draw conclusive implications. Another explanation for the apparent lack of backtracking for time period 1 is that, in line with the experimental hypothesis, fewer participants might extend their causal inferences to the time periods further away from the time period affected by the counterfactual supposition (hypothesis 4). Finally, conditional analyses showed that the majority of participants displayed level 3 causal reasoning: their causal inferences about Night 1 and Night 5 were conditional and consistent with their inferences about Night 2 and Night 4 respectively (this was true for both conditions).

4.6 Experiment 2

Experiment 2 aims to explore how these findings may extend to causal loops involving different factors, or scenarios. Specifically, the scenario in Experiment 1 involved a counterfactual supposition that affected the same factor (amount of sleep) that participants were also asked to make causal inferences on (sleep at past and future time periods). For instance, in the negative loop scenario, sleep affected fatigue, which affected sleep. However, both the counterfactual manipulation and the causal inferences were based on the amount of sleep rather than the amount of fatigue. This meant that even though participants had to take into account the state of both factors, they had to reason explicitly only about the state of one of the factors.

Experiment 2 will follow a similar format to Experiment 1, but will be based on a scenario involving a predator-prey relationship (only the negative loop version will be explored). A simple case where an increase in the number of preys leads to an

increase in the number of predators, which will lead to a decrease in the number of preys.

Another key reason why the predator-prey relationship was deemed a suitable choice of scenario, is because it is often used in system dynamics studies exploring dynamic thinking (e.g. Dorner & Preubler, 1990). In such studies participants are asked to control a population (e.g. the predators) in order to keep another population (e.g. the prey) at a predefined level. As discussed in the Introduction, these studies form the basis of the conclusion that people are unable to appreciate and reason with causal loops. Hence, it is worth considering if people can indeed reason with the cyclical causal relations of a simple predator-prey causal loop.

Finally, another motive why predator-prey loops are important to consider is because they are a prime example of a situation where people's linear thinking is argued to lead to overexploitation or extinction of natural resources (Moxnes, 2003). Given these potentially dramatic implications, it is critical to elucidate further the reasoning patterns that might be involved.

The experimental hypotheses for Experiment 2 are similar to the ones for Experiment 1:

Hypothesis 1: Participants will backtrack only in the observation condition and not in the intervention condition.

Hypothesis 2: Participants will make the same forward inferences in both conditions.

Hypothesis 3: The answers selected by participants will reflect the correct qualitative causal inference according to the counterfactual supposition.

Hypothesis 4: Fewer participants will extend their causal inferences to the time periods further away from the time period affected by the counterfactual

supposition.

Hypothesis 5: Participants' causal inferences about Time 1 and Time 5 will be conditional and consistent with their inferences about Time 2 and Time 4 respectively.

4.7 Method

Participants

99 participants were recruited through Amazon Turk. Participation in the survey was limited to people living in the United States to maximize likelihood of recruiting participants who speak English as their first language. The study was advertised as investigating reasoning about causes and effects. All participants were paid $0.50. 61 were males (61.6%) and 38 were females (38.4%). The mean age was 32.8 years (SD = 11.1; range 19 to 80 years). The participants' education background was approximately equally split between Sciences (N=37; 37.4%), Arts (N=25; 25.3%) and Mixed (N=37; 37.4%).

Design

The experiment comprised two between-subjects conditions: the Observation condition and the Intervention condition. All participants were presented with the same scenario. The dependent variable was the answers to the four questions - a question about each time period. Participants were always questioned about the two past time periods first and the future ones after, but the order in which they were questioned about them was counterbalanced. The survey ended with a series of demographic questions.

Materials

The materials consisted of a web-based questionnaire. The first page of the questionnaire provided simple instructions telling participants they will be presented with one scenario and asked to answer questions about it. Then all participants were presented with the scenario followed by the counterfactual manipulation (either based on an observation or an intervention according to condition). After the manipulation participants were presented with the four causal inference questions. Finally, participants were given a series of demographic questions.

Scenario. The scenario was based on a predator-prey relationship involving tuna (the predator) and squid (the prey). The exact wording of the scenario is reported below.

> Tuna prey on squid. This relationship between tuna and squid has an effect on the population size of both tuna and squid. Tuna eat squid, whose population is therefore decreased. With fewer squid available the tuna are in greater competition with each other for the remaining squid. The tuna population is therefore reduced because some tuna are unable to obtain enough squid for their survival. With fewer tuna left, fewer squid are eaten. The squid population therefore increases. With more squid available, the tuna population in turn increases. This results in cyclical fluctuations in the population sizes of tuna and squid. Imagine that there are a number of tuna and squid in one specific area of the ocean. Assume that the number of squid grows rapidly when tuna are absent. Also assume that the tuna population will starve in the absence of the squid population (as opposed to switching to another type of

prey). The table below shows the number of tuna and squid that are present during five time periods.

Time period	Animal population
Time 1	1000 squid
Time 2	200 tuna
Time 3	500 squid
Time 4	100 tuna
Time 5	1000 squid

Counterfactual manipulations. The exact wording of the suppositions in the two conditions is reported below.

Intervention condition: Suppose at time period 3, contrary to what is stated above, scientists had come to the area of the ocean (described above) for the first time. Imagine they removed 250 of the 500 squid.

Observation condition: Suppose that at time period 3, contrary to what is stated above, there had actually been 250 squid and not 500 squid.

Questions. The four causal inference questions were presented on two separate pages. The first page contained the two questions about the two time periods closest to the time period affected by the counterfactual supposition (Time 2 and Time 4). The second page contained the two questions about the two time periods further away from the time tie period affected by the counterfactual supposition (Time 1 and Time 5). The scenario and counterfactual supposition was displayed at the top of each page so participants could refer back to it. The questions all followed the same multiple-choice format. The exact wording for one of the questions is reported below.

E.g. Time 1: Given this new piece of information about time period 3, what can you infer about the population of squid at time period 1?

- There would have been less than 1000 squid at time period 1.
- There would have been more than 1000 squid at time period 1.
- There would still have been 1000 squid at time period 1.

The order in which the answer choices were presented was randomized.

Procedure

The procedure was as in Experiment 1.

4. 8 Results

Table 7 shows the percentage of participants (and number) in each answer category for the four questions, as a function of condition. The 'correct' normative answer for each question is indicated in bold. The analyses will be discussed in respect to each of the five experimental hypotheses.

Hypothesis 1: Participants will backtrack (change causal inferences according to the counterfactual supposition for both Time 2 and Time 1) only in the observation condition and not in the intervention condition.

As can be observed in Table 7, most participants in the Intervention condition do not backtrack when asked about Time 2 (65.3%) and Time 1 (77.6%) - they state the number of tuna would not have changed as a result of the counterfactual supposition. On the other hand, in the Observation condition, only a few participants answer 'same' when asked about Time 2 (12%). Fewer participants backtrack at Time 1, as 40% answer 'same'. A chi-square for independence was run to compare the

observed frequency of cases in each condition for both Time 2 and Time 1. As predicted by the hypothesis, there was a significant association between condition and answer choice for Time 2: χ^2 (2, n=99) = 34.1 p < 0.001, phi = 0.59; and for Time 1: χ^2 (2, n=99) = 15.1 p = 0.001, phi = 0.39.

Table 7.

Percentage of participants (and number) in each answer category for the four questions, as a function of condition. Values highlighted in bold indicate the correct normative answers for each condition. Experiment 2.

Question	Answer	Intervention (N=49)	Observation (N=50)
Time 1	Less	16.3% (N=8)	**34% (N=17)**
	More	6.1% (N=3)	26% (N=13)
	Same	**77.6% (N=38)**	40% (N=20)
Time 2	Less	22.4% (N=11)	28% (N=14)
	More	12.2% (N=6)	**60% (N=30)**
	Same	**65.3% (N=32)**	12% (N=6)
Time 4	Less	**77.6% (N=38)**	**70% (N=35)**
	More	12.2% (N=6)	18% (N=9)
	Same	10.2% (N=5)	12% (N=6)
Time 5	Less	40.8% (N=20)	44% (N=22)
	More	**38.8% (N=19)**	**36% (N=18)**
	Same	20.4% (N=10)	20% (N=10)

Hypothesis 2: Participants will make the same forward inferences in both conditions (change causal inferences according to the counterfactual supposition for both Time 4 and Time 5).

As can be observed in Table 7, in the Intervention condition only a few participants state the number of tuna would be the same at Time 4 (10.2%) and Time 5 (20.4%). A similar pattern of results is found in the Observation condition for Time 4 (12%) and Time 5 (20%). A chi-square for independence was run to compare the observed frequency of cases in each condition for both Time 4 and Time 5. As predicted by the hypothesis, there was no significant association between condition and answer choice for Time 4: χ^2 (2, n=99) = 0.8, p = 0.669, phi = 0.09; nor for Time 5: χ^2 (2, n=99) = 0.11, p = 0.945, phi = 0.034.

Hypothesis 3: The answers selected by participants will reflect the correct qualitative causal inference according to the counterfactual supposition.

As can be observed in Table 7, the correct predicted normative answer is indicated in bold and the majority of participants select that answer for Time 2 and Time 4 in both conditions, as well as for Time 1 in the Intervention condition. All these differences are significantly greater than chance (binomial test, $p<0.001$). On the other hand, this is not true for Time 1 in the Observation condition (answers split between 'less' and 'same') and Time 5 in both conditions (answers split between 'less' and 'more').

Hypothesis 4: Fewer participants will extend their causal inferences to Time 1 and Time 5.

In the Observation condition the majority of participants selects the correct normative answer for Time 2 (60% select 'more'), but not for Time 1 (only 34% select 'less'). Participants' answers were recoded into two categories: correct and incorrect. These were entered in a chi-square test for independence comparing Time 1

to Time 2. The chi-square revealed a significant difference between the two time periods: χ^2= (1, n=50) 14.08, $p < 0.001$. The same can be said about Time 5. Even though in the Observation condition the majority of participants selects the correct normative answer for Time 4 (70% answer 'less'), remarkably fewer participants do so for Time 5 (36% select 'more'). This difference is significant: $\chi^2(1, n=50) = 1333$, $p < 0.001$.

When it comes to the Intervention condition, contrary to what was predicted, a smaller number of people select the right answer for Time 2 (65.3% select 'same') than for Time 1 (77.6% selects 'same'). However, this difference is not significant: $\chi^2(1, n=549) = 3.24$, $p =0.07$.On the other hand, when it comes to inferences about the future, a greater number selects the correct answer for Night 4 (77.6% select 'less') than for Night 5 (38.8% select 'more). This difference is significant: $\chi^2(1, n=49) = 42.32$, $p = 0$.

Hypothesis 5: Participants' causal inferences about Time 1 and Time 5 will be conditional and consistent with their inferences about Time 2 and Time 4 respectively.

Causal inferences about Time 1 should be contingent on causal inferences about Time 2. Participants' answers for Time 1 given their answer for Time 2 are displayed in Table 8. As can be seen in Table 8, contrary to the hypothesis, for the Observation condition only a minority of participants who correctly infer that there would be more tuna at Time 2 also correctly infer that there would less tuna at Time 1 (N=20%) – the responses are more or less equally split between the three answers.

For the Intervention condition, the majority of participants who correctly infer that there would be the same number of tuna at Time 2 also correctly infer that there would be the same number of tuna at Time 1 (55.1%).

Table 8.

Table showing percentage of participants' answers to Time 1 as a function of their answer to Time 2. Values highlighted in bold indicate the correct normative answers for each condition. Experiment 2.

Answer for Time 2	Answer for Time 1	Intervention	Observation
Less	Less	2.0% (N=1)	12.0% (N=6)
	More	0.0% (N=0)	8.0% (N=4)
	Same	20.4% (N=10)	8.0% (N=4)
More	Less	6.1% (N=3)	**20.0% (N=10)**
	More	4.1% (N=2)	18.0% (N=9)
	Same	2.0% (N=1)	22.0% (N=11)
Same	Less	8.2% (N=4)	2.0% (N=1)
	More	2.0% (N=1)	0.0% (N=0)
	Same	**55.1% (N=27)**	10.0% (N=5)

Similarly, causal inferences about Time 5 should be contingent on causal inferences about Time 4. Participants' answers for Time 5 given their answer for Time 4 are displayed in Table 9. As can be seen in Table 9 and in accordance to to the hypothesis, for both conditions the majority participants who correctly infer that there would be less tuna at Time 4 also correctly infer there would be more Tuna at 5 (Intervention condition: N = 34.7%; Observation condition: 30%).

Table 9.

Table showing percentage of participants' answers to Time 5 as a function of their answer to Time 4. Values highlighted in bold indicate the correct normative answers for each condition. Experiment 2.

Answer for Time 4	Answer for Time 5	Intervention	Observation
Less	Less	26.5% (N=13)	28.0% (N=14)
	More	**34.7% (N=17)**	**30.0% (N=15)**
	Same	16.3% (N=8)	12.0% (N=6)
More	Less	4.1% (N=2)	10.0% (N=5)
	More	4.1% (N=2)	6.0% (N=3)
	Same	4.1% (N=2)	2.0% (N=1)
Same	Less	10.2% (N=5)	6.0% (N=3)
	More	0.0% (N=0)	0.0% (N=0)
	Same	0.0% (N=0)	6.0% (N=3)

4. 9 Discussion

Experiment 1 provided support for the hypothesis that people are able to reason with simple causal loops and make sensible causal inferences accordingly. Experiment 2 aimed to extend these findings to a more complex loop. The added complexity was derived namely by the fact that the counterfactual supposition affected a different causal factor (number of preys) than the one participants were asked to make causal inferences on (number of predators). This meant that, in contrast to Experiment 1, participants had to reason explicitly about the state of both causal factors (rather than reasoning explicitly only about the state of only one of the

factors). The findings from Experiment 2 provided further support for the hypothesis that people are able to reason with simple causal loops and make sensible causal inferences accordingly.

Negative loop

As in Experiment 1, participants clearly engaged in level 1 causal reasoning. They backtracked only in the Observation condition and not in the Intervention condition (hypothesis 1). The difference in backtracking between conditions was significant. Secondly, participants made similar forward inferences in both conditions - there was no significant differences between conditions (hypothesis 2). On the other hand, Experiment 2 provided mixed evidence for level 2 causal reasoning. It was predicted that the answers selected by participants would reflect the correct qualitative causal inference according to the counterfactual supposition (hypothesis 3). Participants' answers for Time 2 and 4 were certainly consistent with this prediction (both conditions). However, when it came to Time 1 and 5, participants gave mixed answers. As in Experiment 1, for Time 1 in the Observation condition, 40% of participants answered 'same' indicating a lack of backtracking. Furthermore, for time period 5, in both conditions, the answers were split between 'less' and 'more'. This finding is in line with the hypothesis that fewer participants will extend their causal inferences to the time periods further away from the time period affected by the counterfactual supposition (hypothesis 4). Given that many participants did not attain level 2 causal reasoning, it follows that their inferences cannot be classified as having the coherency required by level 3. Therefore, in contradiction to what was predicted, conditional analyses showed that participants' inferences about Time 1 and Time 5

were not conditional and consistent with their inferences about Time 2 and Time 4 respectively (hypothesis 5).

4.10 Experiment 3

At first glance, given the results obtained for the negative loop scenario, it may be tempting to conclude that participants simply cannot deal with the added complexity of the causal loop scenario employed in Experiment 2. . However, an alternative explanation may reside in the arguably confusing nature of the framing of the experimental manipulations. Specifically, in the Observation condition, participants are told: "Suppose that at time period 3, contrary to what is stated above, there had actually been 250 squid and not 500 squid."

The manipulation does not provide an explanation for this observation. Therefore there is the possibility that participants may interpret such observation as an error in the data (which is, arguably, a form of intervention). Such interpretation, could lead them to assume that there is a systematic problem with the data reported at the different time periods. This assumption could create a source of confusion when making inferences about past and future. This idea is consistent with the results that seem to indicate that there is more confusion in the Observation condition rather than in the Intervention condition.

The aim of Experiment 3, as well as extending the current findings to another scenario, is to attempt to tease apart this potential source of confusion. This can be achieved by creating counterfactual suppositions which do not affect the causal loop actually described in the scenario but, affect another loop with the same properties - which is essentially identical to the one described in the scenario. This loop with identical causal properties will be termed the 'parallel loop'. The effect of this

variation can be explored by comparing the results from the conditions with the current counterfactual suppositions with results from additional conditions where the suppositions affect the parallel loop.

Experiment 3 will comprise four conditions: two standard conditions (Intervention condition and Observation condition) and two additional conditions (Parallel intervention condition and Parallel observation condition). The scenario will be based on a causal loop involving overfishing. In the negative loop version, an increase in a country's overfishing of tunas leads to an increase in the country's conservation efforts (fishing ban), which will lead to a decrease in overfishing. In the positive loop version, an increase in a country's overfishing of tunas leads to an increase in the country's price of fish, which will lead to an increase in overfishing (higher demand).

The two conditions involving the parallel loop will use the same scenario as the standard conditions, but the supposition will not affect the country described in the scenario (Country A); instead it will affect another country (Country B) identical to Country A (see Method for further details). Hence, participants' causal inferences will be based on Country B. This should prevent them for inferring that the supposition described in the Observation condition is a mistake in the data count or any similar systematic error. Rather, it is just different data on a different country.

Another potential source of confusion in Experiment 2, might reside in the fact that participants were told to imagine that the state of the variable affected by the counterfactual supposition (number of squid), would change 'at Time period 3'. Basically the supposition does not specify at which time point (start, middle, end) during the time period the variable would change. This means that, in theory, even if the state of squid changes during time period 3, some participants might assume this

change only affects the state of the other variable (tuna) during time period 4). This could also explain the mixed results. Therefore, to clarify things even further, the state of the other variable (fishing in this case) will be reported in the table as well. In addition the supposition will specify 'at the start of time period 3…'.

Finally, like the predator-prey loop described in Experiment 2, the overfishing cover story was chosen because it is yet another example of a situation where people's linear thinking is argued to lead to overexploitation or extinction of natural resources. The experimental hypotheses are identical to those in Experiment 2, with the added prediction that the new format will improve participants' causal reasoning. The exact hypothesis is reported below.

> Hypothesis 6: Participants in the Parallel conditions will make more correct causal inferences than participants in the Standard conditions.

4.11 Method

Participants

532 participants were recruited through Amazon Turk. Participation in the survey was limited to people living in the United States to maximize likelihood of recruiting participants who speak English as their first language. The study was advertised as investigating reasoning about causes and effects. All participants were paid $0.50. 333 were males (62.6%) and 199 were females (37.4%). The mean age was 29.5 years (SD = 9.6; range 18 to 69 years). The participants' education background was approximately equally split between Sciences (N=201; 37.8%), Arts (N=129; 24.2%) and Mixed (N=202; 39%).

Design

The experiment comprised four between-subjects conditions: the Standard observation condition, the Standard intervention condition, the Parallel observation condition and the Parallel intervention condition. All participants were presented with the same scenarios. The dependent variable was the answers to the four questions - a question about each time period. Participants were always questioned about the two past time periods first and the future ones after, but the order in which they were questioned about them was counterbalanced. The survey ended with a series of demographic questions.

Materials

The materials consisted of a web-based questionnaire. The first page of the questionnaire provided simple instructions telling participants they will be presented with two scenarios and asked to answer questions about them. Then, depending on the counterbalancing condition, participants were either presented either with the positive loop scenario followed by the negative loop scenario, or vice-versa. Following each scenario, participants were presented with the counterfactual manipulation (either based on an observation or an intervention according to condition). After the manipulation participants were presented with the four causal inference questions. Finally, participants were given a series of demographic questions.

Scenarios. The two scenarios were both based on overfishing and were designed to be similar to each other in terms of cover story. However they were not designed to be directly comparable to each other. The exact wording of each scenario is reported below.

Negative loop

In the ocean surrounding one country (Country A) there are a number of tunas. At the start of each month, scientists estimate the current number of tunas in the ocean surrounding Country A. The government of Country A uses the scientists' monthly estimates to regulate the country's fishing policy. When scientists estimate that there are 1000 tunas or more at the start of the month, the government allows fishing of tunas throughout that month. When scientists estimate that there are 800 tunas or less at the start of the month, the government increases conservation efforts and bans fishing of tunas throughout that month. One month of fishing of tunas causes the number of tunas to drop to 800 or less. One month of no fishing of tunas causes the number of tunas to increase again to 1000 or more. The table below shows the estimated number of tunas at the start of five months and the country's fishing policy throughout those months.

Time	Country A
Start of month 1	800 tunas
Month 1	No fishing
Start of month 2	1000 tunas
Month 2	Fishing
Start of month 3	800 tunas
Month 3	No fishing
Start of month 4	1000 tunas
Month 4	Fishing
Start of month 5	800 tunas
Month 5	No fishing

Positive loop

In the ocean surrounding one country (Country A) there are a number of tunas. At the start of each month scientists estimate the current number of tunas in the ocean surrounding Country A. The government of Country A does not regulate fishing of tunas. Fishing of tunas causes the number of tunas to drop by 200 every month. The table below shows the estimated number of tunas at the start of five months and the country's fishing policy throughout those months.

Time	Country A
Start of month 1	1200 tunas
Month 1	Fishing
Start of month 2	1000 tunas
Month 2	Fishing
Start of month 3	800 tunas
Month 3	Fishing
Start of month 4	600 tunas
Month 4	Fishing
Start of month 5	400 tunas
Month 5	Fishing

Counterfactual manipulations

The exact wording of the suppositions in the four conditions is reported below.

Standard intervention condition: Suppose that at the start of month 3, before estimating the current number of tunas, scientists introduced 200 additional

tunas into the ocean (same tunas as the ones already there). Imagine this resulted in the estimated number of tunas at the start of month 3 being 1000 and not 800.

Standard observation condition: Suppose that at the start of month 3, contrary to what is stated above, the estimated number of tunas was 1000 and not 800.

Parallel intervention condition: Suppose there is another country (Country B) identical to the country described above (Country A). Country B operates in the same way as Country A. This means Country B's government regulates the fishing policy according to the scientists' monthly estimates of the current number of tunas. Suppose that in Country B, at the start of month 3, before estimating the current number of tunas, scientists introduced 200 tunas into the ocean (same tunas as the ones already there). Imagine this resulted in the estimated number of tunas at the start of month 3 being 1000 and not 800.

Parallel observation condition: Suppose there is another country (Country B) identical to the country described above (Country A). Country B operates in the same way as Country A. This means Country B's government regulates the fishing policy according to the scientists' monthly estimates of the current number of tunas. Suppose that in Country B, at the start of month 3, the estimated number of tunas was 1000.

Questions. The four causal inference questions were presented on two separate pages. The first page contained the two questions about the two time periods closest to the time period affected by the counterfactual supposition (Time 2 and Time 4). The second page contained the two questions about the two time periods further away from the time period affected by the counterfactual supposition (Time 1 and Time 5). The scenario and counterfactual supposition was displayed at the top of each

page so participants could refer back to it. The questions all followed the same multiple-choice format. The exact wording for one of the questions is reported below.

E.g. Time 1 (Parallel conditions): Given this information about Country B, what can you infer about the number of tunas in Country B at *month 1*?

- There would be less than 800 tunas at month 1.
- There would be more than 800 tunas at month 1.
- There would be 800 tunas at month 1.

The order in which the answer choices were presented was randomized.

Procedure

The procedure was as in Experiments 1 and 2.

4. 12 Results

The results for each scenario will be discussed separately. Results from the parallel conditions will be discussed alongside the results from the standard conditions.

Negative loop

Table 10 and Table 11 show the percentage of participants (and number) in each answer category for the four questions, as a function of condition. Table 10 displays results from the two standard conditions whilst Table 11 displays the ones from the two parallel conditions. In each table, the 'correct' normative answer for each question is indicated in bold.

Table 10.

Percentage of participants (and number) in each answer category for the four questions, as a function of each standard condition. Values highlighted in bold indicate the correct normative answers for each condition. Experiment 3, negative loop scenario.

Question	Answer	Intervention (N=132)	Observation (N=136)
Time 1	Less	3.8% (N=5)	5.9% (N=8)
	More	12.9% (N=17)	**50% (N=68)**
	Same	**83.3% (N=110)**	44.1% (N=60)
Time 2	Less	6.1% (N=8)	**24.3% (N=33)**
	More	5.3% (N=7)	44.9% (N=52)
	Same	**88.6% (N=117)**	30.9% (N=22)
Time 4	Less	**51.5% (N=68)**	**45.6% (N=62)**
	More	38.6% (N=51)	*38.2% (N=52)*
	Same	*9.8% (N=13)*	16.2% (N=22)
Time 5	Less	3% (N=4)	8.8% (N=12)
	More	**88.6% (N=117)**	***66.9% (N=91)***
	Same	*8.3% (N=11)*	24.3% (N=33)

The analyses will be discussed in respect to each of the six experimental hypotheses.

Hypothesis 1: Participants will backtrack (change causal inferences according to the counterfactual supposition for both Time 2 and Time 1) only in the observation conditions (standard and parallel) and not in the intervention conditions (standard and parallel).

As can be observed in Table 10, most participants in the Standard intervention condition do not backtrack when asked about Time 2 (88.6%) or Time 1 (83.3%) -

they state the number of tuna would not have changed according to the counterfactual supposition. On the other hand, in the Standard observation condition, only a few answer 'same' when asked about Time 2 (30.9%) and Time 1 (44.1%). A chi-square for independence was run to compare the observed frequency of cases in each of these two conditions for both Time 2 and Time 1. As predicted by the hypothesis, there was a significant association between condition and answer choice for Time 2: χ^2 (2, n=268) = 93.4, $p < 0.001$, phi = 0.59; and for Time 1: χ^2 (2, n=268) = 45.9 $p < 0.001$, phi = 0.41.

Table 11.

Percentage of participants (and number) in each answer category for the four questions, as a function of each parallel condition. Values highlighted in bold indicate the correct normative answers for each condition. Experiment 3, negative loop scenario.

Question	Answer	Intervention (N=133)	Observation (N=131)
Time 1	Less	7.5% (N=10)	6.1% (N=8)
	More	17.3% (N=23)	**86.3% (N=113)**
	Same	**75.2% (N=100)**	7.6% (N=10)
Time 2	Less	18.8% (N=25)	**59.5% (N=78)**
	More	9.8% (N=13)	28.2% (N=37)
	Same	**71.4% (N=95)**	12.2% (N=16)
Time 4	Less	**57.9% (N=77)**	**67.2% (N=62)**
	More	27.1% (N=36)	18.3% (N=9)
	Same	15% (N=20)	14.5% (N=6)
Time 5	Less	5.3% (N=7)	7.6% (N=10)
	More	**81.2% (N=108)**	**81.7% (N=107)**
	Same	13.5% (N=18)	10.7% (N=14)

Similarly, as can be observed in Table 11, most participants in the Parallel intervention condition do not backtrack when asked about Time 2 (71.4%) and Time 1 (75.2%). On the other hand, in the Parallel observation condition, only a few answer 'same' when asked about Time 2 (12.2%) and Time 1 (7.6%). As predicted by the hypothesis, there was a significant association between condition and answer choice for Time 2: χ^2 (2, n=264) = 95, $p < 0.001$, phi = 0.6; and for Time 1: χ^2 (2, n=264) = 133.4 $p < 0.001$, phi = 0.71.

Hypothesis 2: Participants will make the same forward inferences in all conditions (change causal inferences according to the counterfactual supposition for both Time 4 and Time 5).

As can be observed in Table 10, in the Standard intervention condition only a few participants state the number of tuna would be the same at Time 4 (9.8%) and Time 5 (8.3%). A similar pattern of results can be observed in the Standard observation condition for Time 4 (16.2%) and Time 5 (24.3%). A chi-square for independence was run to compare the observed frequency of cases in these two conditions for both Time 4 and Time 5. As predicted by the hypothesis, there was not a significant association between condition and answer choice for Night 4: χ^2 (2, n=268) 2.54, $p = 0.281$, phi = 0.1. However, the number of people who state there would be the same number of tuna at Time 5 is significantly greater in the Observation condition than in the Intervention condition: χ^2 (2, n=268) = 18.2, $p < 0.001$, phi = 0.26.

Similarly, as can be observed in Table 11, in the Parallel intervention condition only a few participants state the number of tuna would be the same at Time 4 (15%) and Time 5 (13.5%). A similar pattern of results can be observed in the

Parallel observation condition for Time 4 (10.1%) and Time 5 (15.9%). As predicted by the hypothesis, there was not a significant association between condition and answer choice for Night 4: χ^2 (2, n=264) 3.14, p = 0.208, phi = 0.11; nor for Night 5: χ^2 (2, n=264) = 1.02, p = 0.6, phi = 0.06.

Hypothesis 3: The answers selected by participants will reflect the correct qualitative causal inference according to the counterfactual supposition.

As can be observed in Table 10, the correct predicted normative answer is indicated in bold. For the Standard intervention condition, the majority of participants select that answer for each time period. In all cases, the deviation from 0.3 was highly significant (binomial test, p<0.001). When it comes to the Standard observation condition, the majority of participants select the correct answer from Time 1, 4 and 5. However, this is significantly higher than chance only for Time 1 (binomial test, p=0.002) and Time 5 (binomial test, p<0.001); not for Time 4 (p=0.086). Unexpectedly, for Time 2, less than a third of participants make the correct choice (N=24.3%; binomial test, p=0.558)

On the other hand, for the Parallel conditions, as can be observed in Table 11, the correct predicted normative answer is indicated in bold and the majority of participants select that answer for each time period and for both conditions. The difference was significant for all time periods (Time periods 1, 2 and 5: binomial test, p<0.001; Time 4: p=0.029).

Hypothesis 4: Fewer participants will extend their causal inferences to Time 1 and Time 5.

When it comes to the Standard intervention condition, in line with the

experimental hypothesis, slightly more participants select the correct answer for Time 2 (88.6% select 'same') than for Time 1 (83.3% selects 'same'). Participants' answers were recoded into two categories: correct and incorrect. These were entered in a chi-square test for independence comparing Time 1 to Time 2. The chi-square revealed a significant difference between the two time periods: χ^2= (1, n=132) 3.685, p = 0.005. This does not stand for Time 5 in the Standard intervention condition: contrary to what was predicted, a considerably smaller number of people select the correct answer for Time 4 (51.5% select 'less) than for Time 5 (88.6% select 'more'). The chi-square revealed a significant difference between the two time periods: χ^2= (1, n=132) 72.8, p < 0.001.

Again, when it comes to the Standard observation condition, contrary to what was predicted, more participants give correct answers for Time 1 (50% select 'more') than for Time 2 (44.9% 'less'). This difference is significant: χ^2(1, n=136) = 49, p <0.001. Similarly, more participants give correct answers for Time 5 (66.9% select 'more') than for Time 4 (45.6% 'less'). This difference is significant: χ^2(1, n=136) = 24.9, p =0.001.

In the Parallel intervention condition, contrary to what was predicted, a smaller number of people select the correct answer for Time 2 (71.4% select 'same') than for Time 1 (75.2% select 'same'). However this difference is not significant: χ^2= (1, n=133) 0.921, p =0.33. Similarly, a smaller number of people select the correct answer for Time 4 (57.9% select 'less') than for Time 5 (81.2% select 'more). This difference is significant: χ^2= (1, n=133) 29.6, p <0.001.

When it comes to the Parallel observation condition, contrary to what was predicted, a considerably smaller number of people select the correct answer for Time 2 (59.5% select 'less') than for Time 1 (86.3% select 'more'). This difference is

significant: $\chi^2(1, n=131) = 38.8$, $p < 0.001$. The same can be said about Time 5: contrary to what was predicted, a considerably smaller number of people select the correct answer for Time 4 (67.2% select 'less') than for Time 5 (81.7% select 'more'). This difference is significant: $\chi^2(1, n=131) = 62$, $p < 0.001$.

Hypothesis 5: Participants' causal inferences about Time 1 and Time 5 will be conditional and consistent with their inferences about Time 2 and Time 4 respectively.

Table 12 shows participants' answers for Time 1 given their answer for Time 2, for the two Standard conditions. As can be seen in Table 12, for the Standard Observation condition, the majority of participants infer (wrongly) that there would be more tuna at Time 2 and more tuna at Time 1 (27.2%). This means that fewer participants give consistent correct normative answers; i.e. indicate there would less tuna at Time 2 and then more tuna at Time 1 (16.2%). For the Standard intervention condition, the majority of participants who correctly infer that there will be the same number of tuna at Time 2, also infer that there would be the same number of tuna at Time 1 (N=80.3%).

Participants' answers for Time 5 given their answer for Time 4 are displayed in Table 13. As can be seen in Table 13, for both conditions, the majority of participants who correctly infer that there would be less tuna at Time 4 infer that there would be more tuna at Time 5 (Intervention condition: N = 48.5%; Observation condition: 34.6%).

Table 12.

Table showing percentage of participants' answers to Time 1 as a function of their answer to Time 2. Values highlighted in bold indicate the correct normative answers for each condition. Experiment 3, negative loop scenario.

Answer for Time 2	Answer for Time 1	Intervention	Observation
Less	Less	0.0% (N=0)	2.2% (N=3)
	More	5.3% (N=7)	**16.2% (N=22)**
	Same	0.8% (N=1)	5.9% (N=8)
More	Less	1.5% (N=2)	1.5% (N=2)
	More	1.5% (N=2)	27.2% (N=37)
	Same	2.3% (N=3)	16.2% (N=22)
Same	Less	2.3% (N=3)	2.2% (N=3)
	More	6.1% (N=8)	6.6% (N=9)
	Same	**80.3% (N=106)**	22.1% (N=30)

Table 14 shows participants' answers for Time 1 given their answer for Time 2, for the two Parallel conditions. As can be seen in Table 14, for the Parallel observation condition, the majority of participants who correctly infer that there would be less tuna at Time 2 also infer that there would be more tuna at Time 1 (N=55.1%).

For the Parallel intervention condition, the majority participants who correctly infer that there will be the same number of tuna at Time 2, also infer that there would be the same number of tuna at Time 1 (N=62.1%).

Table 13.

Table showing percentage of participants' answers to Time 5 as a function of their answer to Time 4. Values highlighted in bold indicate the correct normative answers. Experiment 3, negative loop scenario.

Answer for Time 4	Answer for Time 5	Intervention	Observation
Less	Less	2.3% (N=3)	4.4% (N=6)
	More	**48.5% (N=64)**	**34.6% (N=47)**
	Same	0.8% (N=1)	6.6% (N=9)
More	Less	0.0% (N=0)	4.4% (N=6)
	More	35.6% (N=47)	27.2% (N=37)
	Same	3.0% (N=4)	6.6% (N=9)
Same	Less	0.8% (N=1)	0.0% (N=0)
	More	4.5% (N=6)	5.1% (N=7)
	Same	4.5% (N=6)	11.0% (N=15)

Table 14.

Table showing percentage of participants' answers to Time 1 as a function of their answer to Time 2. Values highlighted in bold indicate the correct normative answers. Experiment 3, negative loop scenario.

Answer for Time 2	Answer for Time 1	Intervention	Observation
Less	Less	2.3% (N=3)	0.7% (N=1)
	More	6.1% (N=8)	**55.1% (N=75)**
	Same	10.6% (N=14)	1.5% (N=2)
More	Less	3.8% (N=5)	2.9% (N=4)
	More	3.0% (N=4)	22.1% (N=30)
	Same	3.0% (N=4)	2.2% (N=3)
Same	Less	1.5% (N=2)	2.2% (N=3)
	More	8.3% (N=11)	5.9% (N=8)
	Same	**62.1% (N=82)**	3.7% (N=5)

Participants' answers for Time 5 given their answer for Time 4 are displayed in Table 15. As can be seen in Table 15, for both conditions, the majority of participants who correctly infer that there would be less tuna at Time 4 also infer that there would be more tuna at Time 5 (Intervention condition: N = 47.7%; Observation condition: 55.1%).

Table 15.

Table showing percentage of participants' answers to Time 5 as a function of their answer to Time 4. Values highlighted in bold indicate the correct normative answers for each condition. Experiment 3, negative loop scenario.

Answer for Time 4	Answer for Time 5	Intervention	Observation
Less	Less	3.0% (N=4)	5.1% (N=7)
	More	**47.7% (N=63)**	**55.1% (N=75)**
	Same	7.6% (N=10)	4.4% (N=6)
More	Less	1.5% (N=2)	1.5% (N=2)
	More	22.7% (N=30)	14.0% (N=19)
	Same	3.0% (N=4)	9.6% (N=13)
Same	Less	0.8% (N=1)	0.7% (N=1)
	More	11.4% (N=15)	9.6% (N=13)
	Same	3.0% (N=4)	3.7% (N=5)

Hypothesis 6: Participants in the Parallel conditions will make more correct causal inferences than participants in the Standard conditions.

Table 16 shows the proportion of correct answers for each time period as a function of conditions. Values highlighted in bold indicate the highest proportion of correct answers between the conditions for each time period. Participants' answers were recoded into two categories: correct and incorrect. These were entered in a chi-square test for independence comparing the Standard condition with the Parallel condition (results reported in Table 16). In line with the experimental hypothesis, the Parallel observation condition significantly outperformed the Standard Observation condition for each time period. On the other hand, for the Intervention conditions, it

appears as though the Standard version outperformed the Parallel version for time periods 1, 2 and 5. However, other than for Time 2, the differences are not significant.

Table 16.

Percentage of participants who gave a correct answer for the four questions, as a function of condition. Experiment 3, negative loop scenario.

Question	Condition	Standard condition	Parallel condition	Chi-square value	p-value
Time 1	Intervention	**83.30%**	75.20%	3.518	0.060
	Observation	50%	**86.30%**	111.5	0.000
Time 2	Intervention	**88.60%**	71.40%	14.49	0.000
	Observation	24.30%	**59.50%**	51.42	0.000
Time 4	Intervention	51.50%	**57.90%**	1.4	0.236
	Observation	45.60%	**67.20%**	21.17	0.000
Time 5	Intervention	**88.60%**	81.20%	3.59	0.060
	Observation	66.90%	**81.70%**	14.65	0.000

Positive loop scenario

Table 17 and Table 18 show the percentage of participants (and number) in each answer category for the four questions, as a function of condition. Table 17 displays results from the two standard conditions whilst Table 18 displays the ones from the two parallel conditions. In each table, the 'correct' normative answer for each question is indicated in bold.

Table 17.

Percentage of participants (and number) in each answer category for the four questions, as a function of each standard condition. Values highlighted in bold indicate the correct normative answers for each condition. Experiment 3, positive loop scenario.

Question	Answer	Intervention (N=132)	Observation (N=136)
Time 1	Less	2.3%(N=3)	10.3% (N=14)
	More	4.5% (N=6)	**70.6% (N=96)**
	Same	**93.2% (N=123)**	19.1% (N=26)
Time 2	Less	6.1% (N=8)	5.1% (N=7)
	More	7.6% (N=10)	**71.3% (N=97)**
	Same	**86.4% (N=114)**	23.5% (N=32)
Time 4	Less	1.5% (N=2)	4.4% (N=6)
	More	**91.7% (N=121)**	***85.3% (N=116)***
	Same	*6.8% (N=9)*	10.3% (N=14)
Time 5	Less	3% (N=4)	4.4% (N=6)
	More	**90.9% (N=120)**	**86.8% (N=118)**
	Same	*6.1% (N=8)*	*8.8% (N=12)*

\

Table 18.

Percentage of participants (and number) in each answer category for the four questions, as a function of each parallel condition. Values highlighted in bold indicate the correct normative answers for each condition. Experiment 3, positive loop scenario.

Question	Answer	Intervention (N=133)	Observation (N=131)
Time 1	Less	9.8%(N=13)	4.6% (N=6)
	More	15% (N=20)	**80.2% (N=105)**
	Same	**75.2% (N=100)**	15.3% (N=20)
Time 2	Less	12.8% (N=17)	7.6% (N=10)
	More	12.8% (N=17)	**80.9% (N=106)**
	Same	**74.4% (N=99)**	11.5% (N=15)
Time 4	Less	4.5% (N=6)	6.9% (N=9)
	More	**86.5% (N=115)**	**82.4% (N=108)**
	Same	9% (N=12)	10.7% (N=14)
Time 5	Less	4.5% (N=6)	3.8% (N=5)
	More	**85% (N=113)**	**87.8% (N=115)**
	Same	10.5% (N=14)	8.4% (N=11)

The analyses will be discussed in respect to each of the six experimental hypotheses.

Hypothesis 1: Participants will backtrack (change causal inferences according to the counterfactual supposition for both Time 2 and Time 1) only in the observation conditions (standard and parallel) and not in the intervention conditions (standard and parallel).

As can be observed in Table 17, most participants in the Standard intervention condition do not backtrack when asked about Time 2 (86.4%) and Time 1 (93.2%) - they state the number of tuna would not have changed according to the counterfactual supposition. On the other hand, in the Standard observation condition, only a few answer 'same' when asked about Time 2 (23.5%) and Time 1 (19.1%). A chi-square for independence was run to compare the observed frequency of cases in each of these two conditions for both Time 2 and Time 1. As predicted by the hypothesis, there was a significant association between condition and answer choice for Time 2: χ^2 (2, n=268) = 116.8 $p < 0.001$, phi = 0.66; and for Time 1: χ^2 (2, n=268) = 149.7 $p < 0.001$, phi = 0.75.

Similarly, as can be observed in Table 18, most participants in the Parallel intervention condition do not backtrack when asked about Time 2 (74.4%) and Time 1 (75.2%). On the other hand, in the Parallel observation condition, only a few answer 'same' when asked about Time 2 (11.5%) and Time 1 (15.3%). As predicted by the hypothesis, there was a significant association between condition and answer choice for Time 2: χ^2 (2, n=264) = 128.1, $p < 0.001$, phi = 0.67; and for Time 1: χ^2 (2, n=264) = 113.7 $p < 0.001$, phi = 0.66.

Hypothesis 2: Participants will make the same forward inferences in all conditions (change causal inferences according to the counterfactual supposition for both Time 4 and Time 5).

As can be observed in Table 17, in the Standard intervention condition only a few participants state the number of tuna would be the same at Time 4 (6.8%) and Time 5 (6.1%). A similar pattern of results can be observed in the Standard observation condition for Time 4 (10.3%) and Time 5 (8.8%). A chi-square for

independence was run to compare the observed frequency of cases in these two conditions for both Time 4 and Time 5. As predicted by the hypothesis, there was not a significant association between condition and answer choice for Time 4: χ^2 (2, n=268) 3.12, p = 0.209, phi = 0.1; nor for Time 5: χ^2 (2, n=268) = 1.157, p =561, phi = 0.066.

Similarly, as can be observed in Table 18, in the Parallel intervention condition only a few participants state the number of tuna would be the same at Time 4 (4.5%) and Time 5 (10.5%). A similar pattern of results can be observed in the Parallel observation condition for Time 4 (10.7%) and Time 5 (8.4%). As predicted by the hypothesis, there was not a significant association between condition and answer choice for Time 4: χ^2 (2, n=264) 0.958, p = 0.619, phi = 0.06; nor for Time 5: χ^2 (2, n=264) = 0.453, p = 0.797, phi = 0.04.

Hypothesis 3: The answers selected by participants will reflect the correct qualitative causal inference according to the counterfactual supposition.

As can be observed in Table 17 and 18, the correct predicted normative answer is indicated in bold and the majority of participants select that answer for each time period and for all conditions. In every case, the deviation from 0.3 was highly significant (binomial test, $p<0.001$).

Hypothesis 4: Fewer participants will extend their causal inferences to Time 1 and Time 5.

When it comes to the Standard intervention condition, contrary to the experimental hypothesis, slightly more participants select the correct answer for Time 1 (93.2% select 'same') than for Time 2 (86.4% selects 'same'). This difference is

significant: χ^2= (1, n=132) 5.211, p = 0.022. On the other hand, for Time 5 in the Standard intervention condition, a very similar number of participants select the correct answer for Time 4 (91.7% select 'more') and Time 5 (90.9 % select 'more'): χ^2= (1, n=132) 0.099 p =0.75.

Similarly, when it comes to the Standard observation condition, a very similar number of participants select the correct answer for Time 1 (70.6% select 'more') and Time 2 (71.3% 'more'): χ^2(1, n=132) = 0.039, p =0.844. Likewise, a very similar number of participants select the correct answer for Time 5 (86.8% select 'more') and Time 4 (85.3% 'more'): χ^2(1, n=132) = 0.284, p =0.59.

In the Parallel intervention condition, a very similar number of participants select the correct answer for Time 1 (75.2% select 'same') and Time 2 (74.4% select 'same'): χ^2= (1, n=133) 0.003, p =0.8. Likewise, a very similar number of participants select the correct answer for Time 4 (86.5% select 'more') and Time 5 (85% 'more'): χ^2(1, n=133) = 0.03, p =0.85.

The same can be said for Parallel observation condition, a very similar number of participants select the correct answer for Time 1 (80.2% select 'more') and for Time 2 (80.9% select 'more'): χ^2(1, n=131) = 0.027, p = 0.87. Likewise, a very similar number of participants select the correct answer for Time 5 (87.8% select 'more') and Time 4 (82.4% 'more'): χ^2(1, n=131) = 0.118, p =0.73.

Hypothesis 5: Participants' causal inferences about Time 1 and Time 5 will be conditional and consistent with their inferences about Time 2 and Time 4 respectively.

Table 19 shows participants' answers for Time 1 given their answer for Time 2, for the two Standard conditions. As can be seen in Table 19, for the Standard Observation condition, the majority of participants who correctly infer that there

would be more tuna at Time 2 also infer that there would be more tuna at Time 1 (83.3%). For the Standard observation condition, the majority of participants who correctly infer that there will be the same number of tuna at Time 2, also infer that there would be the same number of tuna at Time 1 (N=61%).

Table 19.

Table showing percentage of participants' answers to Time 1 as a function of their answer to Time 2. Values highlighted in bold indicate the correct normative answers for each condition. Experiment 3, positive loop scenario.

Answer for Time 2	Answer for Time 1	Intervention	Observation
Less	Less	0% (N=0)	2.9% (N=4)
	More	0.8% (N=1)	2.2% (N=3)
	Same	5.3% (N=7)	0% (N=0)
More	Less	0.8% (N=1)	3.7% (N=5)
	More	2.3% (N=3)	**61% (N=83)**
	Same	4.5% (N=6)	6.6% (N=9)
Same	Less	1.5% (N=2)	3.7% (N=5)
	More	3.5% (N=2)	7.4% (N=10)
	Same	**83.3% (N=110)**	5.1% (N=7)

Participants' answers for Time 5 given their answer for Time 4 are displayed in Table 19. As can be seen in Table 20, for both conditions, the majority of participants who correctly infer that there would be more tuna at Time 4 infer that there would be more tuna at Time 5 (Intervention condition: N = 88.6%; Observation condition: 80.1%).

Table 20.

Table showing percentage of participants' answers to Time 5 as a function of their answer to Time 4. Values highlighted in bold indicate the correct normative answers for each condition. Experiment 3, positive loop scenario.

Answer for Time 4	Answer for Time 5	Intervention	Observation
Less	Less	0.7% (N=1)	0.7%(N=1)
	More	0% (N=0)	2.2% (N=3)
	Same	0.8% (N=1)	1.5% (N=2)
More	Less	0.8% (N=1)	2.9% (N=4)
	More	**88.6% (N=117)**	**80.1% (N=109)**
	Same	2.3% (N=3)	2.2% (N=3)
Same	Less	1.5%(N=2)	0.7% (N=1)
	More	2.3% (N=3)	4.4% (N=6)
	Same	3% (N=4)	5.1% (N=7)

Table 21 shows participants' answers for Time 1 given their answer for Time 2, for the two Parallel conditions. As can be seen in Table 21, for the Parallel Observation condition, the majority of participants who correctly infer that there would be more tuna at Time 2 also infer that there would be more tuna at Time 1 (71%). For the Standard observation condition, the majority of participants who correctly infer that there will be the same number of tuna at Time 2, also infer that there would be the same number of tuna at Time 1 (N=67.7%).

Table 21.

Table showing percentage of participants' answers to Time 1 as a function of their answer to Time 2. Values highlighted in bold indicate the correct normative answers for each condition. Experiment 3, positive loop scenario.

Answer for Time 2	Answer for Time 1	Intervention	Observation
Less	Less	6% (N=8)	1.5% (N=2)
	More	3% (N=4)	4.6% (N=6)
	Same	3.8% (N=5)	1.5% (N=2)
More	Less	2.3% (N=3)	3.1% (N=4)
	More	6.8% (N=9)	**71% (N=93)**
	Same	3.8% (N=5)	6.9% (N=9)
Same	Less	1.5% (N=2)	0% (N=0)
	More	5.3% (N=7)	4.6% (N=6)
	Same	**67.7% (N=90)**	6.9% (N=9)

Similarly, causal inferences about Time 5 should be contingent on causal inferences about Time 4. Participants' answers for Time 5 given their answer for Time 4 are displayed in Table 22. As can be seen in Table 22, for both conditions, the majority of participants who correctly infer that there would be more tuna at Time 4 infer that there would be more tuna at Time 5 (Intervention condition: N = 78.9%; Observation condition: 78.6%).

Table 22.

Table showing percentage of participants' answers to Time 5 as a function of their answer to Time 4. Values highlighted in bold indicate the correct normative answers for each condition. Experiment 3, positive loop scenario.

Answer for Time 4	Answer for Time 5	Intervention	Observation
Less	Less	0.8% (N=1)	1.5% (N=2)
	More	1.5% (N=2)	3.1% (N=4)
	Same	2.3% (N=3)	2.3% (N=3)
More	Less	0.8% (N=1)	1.5% (N=2)
	More	**78.9% (N=105)**	**78.6% (N=103)**
	Same	6.9% (N=9)	2.3% (N=3)
Same	Less	3% (N=4)	0.8% (N=1)
	More	4.5% (N=6)	6.1% (N=8)
	Same	1.5% (N=2)	3.8% (N=5)

Hypothesis 6: Participants in the Parallel conditions will make more correct causal inferences than participants in the Standard conditions.

Table 23 shows the proportion of correct answers for each time period as a function of conditions. Values highlighted in bold indicate the highest proportion of correct answers between the conditions for each time period. Participants' answers were recoded into two categories: correct and incorrect. These were entered in a chi-square test for independence comparing the Standard condition with the Parallel condition (results reported in Table 23). The Parallel observation condition significantly outperformed the Standard Observation condition for time periods 1 and 2 (time periods 4 and 5 are more or less equal). On the other hand, for the Intervention

conditions, the Standard version outperformed the Parallel version for all time periods (significantly for time periods 1 and 2).

Table 23.

Percentage of participants who gave a correct answer for the four questions, as a function of condition. Experiment 3, positive loop scenario.

Question	Condition	Standard condition	Parallel condition	Chi-square value	p-value
Time 1	Intervention	**93.2%**	75.2%	17.37	0.001
	Observation	70.6%	**80.2%**	5.8	0.02
Time 2	Intervention	**86.4%**	74.4%	7.56	0.006
	Observation	71.3%	**80.9%**	5.96	0.015
Time 4	Intervention	**91.7%**	86.5%	2.316	0.128
	Observation	**85.3%**	82.4%	0.58	0.446
Time 5	Intervention	**90.9%**	85.0%	2.73	0.098
	Observation	86.8%	**87.8%**	0.093	0.75

4.13 Discussion

Experiment 1 provided support for the hypothesis that people are able to reason with simple causal loops and make sensible causal inferences accordingly. Experiment 2 extended these findings to a more complex loop where participants had to reason explicitly about the state of both causal factors involved in the loop underlying the scenario. In Experiment 1 participants' inferences were clearly consistent with level 3 causal reasoning for both negative and positive loop scenarios.

In Experiment 2 however, participants' answers for the negative loop scenario were noisier - hence it provided mixed support for level 2 causal reasoning. This was true particularly for the Observation condition in the negative loops scenario. Therefore, the aim of Experiment 3 (besides extending findings to another scenario) was to investigate if the noisier responses resulted from the potentially unclear nature of the counterfactual supposition in the Observation conditions (see Discussion of Experiment 2 for further details).

Negative loop scenario

As in the previous experiments, participants attained level 1 causal reasoning. They backtracked only in the Observation conditions and not in the Intervention conditions (hypothesis 1). The difference in backtracking between conditions was significant for both sets of conditions (Standard conditions and Parallel conditions). Secondly, it was predicted that participants would make similar forward inferences in both conditions (hypothesis 2). This was true for all time periods except for Time period 5 in the Standard conditions – results differed slightly between the Intervention and the Observation conditions.

Experiment 3 attempted to increase level 2 causal reasoning by introducing the Parallel conditions. As in the previous experiments, it was predicted that the majority of participants would provide a correct normative answer in both Standard and Parallel conditions (hypothesis 3). For the Standard intervention condition, the majority of participants selected that answer for each time period (significantly higher than chance for all time periods). For the Standard observation condition, the majority of participants selected the correct answer only for Time 1, 4 and 5 (significantly higher than chance only for Time 1 and 5). Unexpectedly, for Time 2, less than third

of participants made the correct choice. On the other hand, for the Parallel conditions, the majority of participants selected the correct answer for each time period and for both conditions (significantly higher than chance for all time periods).

These results provide clear support for the hypothesis that the number of correct answers would be greater in the Parallel conditions (hypothesis 6). In terms of direct comparison of number of correct answers between the two Intervention conditions, for the negative loop scenario there was no important difference between the two at any time period. The number of correct answers was indeed higher for Time period 2 in the Standard intervention condition than in the Parallel intervention condition (in contradiction with the hypothesis) but they were still very high in both sets of conditions. Interestingly, the number of correct answers for time period 4 is low for both sets of intervention conditions (Standard: 51.5%; Parallel 57.9%). That being said, the number of correct responses is still significantly higher than chance in both cases. Therefore these findings support that the idea that the Parallel version of the conditions clarifies the scenario and the supposition.

The second question concerning level 2 causal reasoning is the extent to which inferences are extended to Time 1 and 5 (hypothesis 4). Contrary to the experimental hypothesis, for both sets of conditions, overall participants give more correct answers for times 1 and 5 rather than for times 2 and 4. The only exception is the Standard Intervention condition where participants give significantly more correct answers for Time 2 rather than Time 1. Similarly, in the Parallel intervention condition, even though slightly more participants give correct answers for Time 1, the difference is not significant. On the other hand, when it comes to forward inferences, participants give significantly more correct answers for Time 5 rather than for Time 4 (for both sets of conditions).

Given that participants did not consistently attain level 2 causal reasoning for the two Standard conditions, it follows that their inferences did not have the coherency required by level 3. The problem arises when backtracking in the Observation condition – as mentioned above, the majority selects the wrong normative answer for Time 2. This majority then gives the correct normative answer for Time 1 but this results in an incoherent causal model of the loop. In other words, the conditional analyses show that the majority indicated that the number of tuna increases both at Time 2 and at Time 1. On the other hand, forward inferences are consistent for both conditions. As expected, participants in the parallel conditions displayed level 3 causal reasoning: inferences about Night 1 and Night 5 were conditional and consistent with their inferences about Night 2 and Night 4 respectively (this was true for both conditions in both scenarios).

Positive loop scenario

In contrast to the negative loop scenario, the results were in line with all the experimental hypotheses. This is likely to be due to the fact that the positive loop scenario might be easier to understand and reason with because the changes only happen in one direction. Surprisingly, however, when it comes to comparing the number of correct answers between the two sets of conditions (Standard conditions versus Parallel conditions) the number of correct answers is significantly higher for the Standard version of the Intervention condition for time periods 1 and 2. The reverse is true for the Observation condition – number of correct answers is significantly higher for the Parallel version for both time periods 1 and 2.

4. 14 Experiment 4

Experiment 4 aims to further clarify the counterfactual supposition from the preceding experiments. The supposition used in the Experiments 1 and 2 does not specify at which time point (start, middle, end) during time period 3 the affected variable would change. Therefore Experiment 3 specified 'at the start of time period 3…' However, what Experiment 3 does not specify is at which time point the other variable is then affected by the change in the state of the variable affected by the supposition. In other words, participants are told that the number of tuna has changed at the start of time period 3, but they are not told explicitly at which point in time this affects fishing. This means that participants could assume either that fishing changes (resumes) during time period 3 or during time period 4. In order to resolve this potential source of confusion, Experiment 4 added a specification to the counterfactual supposition: "Therefore suppose that there has been fishing throughout month 3".

4.15 Method

Participants

628 participants were recruited through Amazon Turk. Participation in the survey was limited to people living in the United States to maximize likelihood of recruiting participants who speak English as their first language. The study was advertised as investigating reasoning about causes and effects. All participants were paid $0.50. 391 were males (62.3%) and 237 were females (37.7%). The mean age was 29.2 years (SD = 10.01; range 18 to 72 years). The participants' education background was approximately equally split between Sciences (N=246; 39.2%), Arts (N=167; 26.6%) and Mixed (N=215; 34.2%).

Design

The experiment comprised two between-subjects conditions: the parallel observation condition and the parallel intervention condition. All participants were presented with the same scenario. The dependent variable was the answers to the four questions - a question about each time period. Participants were always questioned about the two past time periods first and the future ones after, but the order in which they were questioned about them was counterbalanced. The survey ended with a series of demographic questions.

Materials

The materials were identical to the ones in Experiment 3 (but only the Parallel conditions and the negative loop scenario were investigated).

Scenario. The scenarios were identical to the two scenarios presented in Experiment 3.

Counterfactual manipulations. The exact wording of the suppositions in the two conditions was identical to the ones in Experiment 3, with the exception that the following sentence was added at the end: "Therefore suppose that there has been fishing throughout month 3".

Questions. The questions were identical to the ones in Experiment 3.

Procedure

The procedure was as in previous experiments.

4.16 Results

Negative loop scenario

Table 24 shows the percentage of participants (and number) in each answer category for the four questions, as a function of condition. The 'correct' normative answer for each question is indicated in bold.

Table 24.

Percentage of participants (and number) in each answer category for the four questions, as a function of condition. Values highlighted in bold indicate the correct normative answers for each condition. Experiment 4, negative loop scenario.

Question	Answer	Intervention (N=321)	Observation (N=307)
Time 1	Less	10.6% (N=34)	5.9% (N=18)
	More	17.4% (N=56)	**82.7% (N=254)**
	Same	**72% (N=231)**	11.4% (N=35)
Time 2	Less	19.3% (N=62)	**72.6% (N=223)**
	More	10.9% (N=35)	18.2% (N=56)
	Same	**69.8% (N=224)**	9.1% (N=28)
Time 4	Less	**50.8% (N=163)**	**82.7% (N=254)**
	More	37.1% (N=119)	7.2% (N=22)
	Same	12.1% (N=39)	10.1% (N=31)
Time 5	Less	9.3% (N=30)	11.4% (N=35)
	More	**77.3% (N=248)**	**76.9% (N=235)**
	Same	13.4% (N=43)	12.7% (N=37)

The analyses will be discussed in respect to each of the five experimental hypotheses.

Hypothesis 1: Participants will backtrack (change causal inferences according to the counterfactual supposition for both Time 2 and Time 1) only in the observation condition and not in the intervention condition.

As can be observed in Table 24, most participants in the Intervention condition do not backtrack when asked about Time 2 (69.8%) and Time 1 (72%) - they state the number of tuna would not have changed according to the counterfactual supposition. On the other hand, in the Observation condition, only a few answer 'same' when asked about Time 2 (9.1%) and Time 1 (11.4%). A chi-square for independence was run to compare the observed frequency of cases in each condition for both Time 2 and Time 1. As predicted by the hypothesis, there was a significant association between condition and answer choice for Time 2: χ^2 (2, n=628) = 248.1 p < 0.001, phi = 0.63; and for Time 1: χ^2 (2, n=628) = 275.6, p = 0.001, phi = 0.66.

Hypothesis 2: Participants will make the same forward inferences in both conditions (change causal inferences according to the counterfactual supposition for both Time 4 and Time 5).

As can be observed in Table 24, in the Intervention condition only a few participants state the number of tuna would be the same at Time 4 (12.1%) and Time 5 (13.4%). A similar pattern of results is found in the Observation condition for Time 4 (10.1%) and Time 5 (12.7%). A chi-square for independence was run to compare the observed frequency of cases in each condition for both Time 4 and Time 5. Contrary to the hypothesis, there was a significant association between condition and answer choice for Time 4: χ^2 (2, n=628) = 87, p <0.001 phi = 0.37. On the other hand, and in accordance to predictions, there was no difference between conditions for Time 5: χ^2 (2, n=628) = 0.873, p = 0.646, phi = 0.037.

Hypothesis 3: The answers selected by participants will reflect the correct qualitative causal inference according to the counterfactual supposition.

As can be observed in Table 24, the correct predicted normative answer is indicated in bold and the majority of participants select that answer for all time periods and conditions. These differences are significantly greater than chance (binomial test, $p<0.001$) for all time periods except for Time 4 in the Intervention condition: only 50.8% of participants select the correct answer ('less'), which results in a non-significant binomial test ($p=0.412$)

Hypothesis 4: Fewer participants will extend their causal inferences to Time 1 and Time 5.

In the Observation condition, contrary to the hypothesis, slightly more participants select the correct normative answer for Time 1 (82.7% select 'more') than for Time 2 (only 72.6% select 'less'). Participants' answers were recoded into two categories: correct and incorrect. These were entered in a chi-square test for independence comparing Time 1 to Time 2. The chi-square revealed a significant difference between the two time periods: χ^2= (1, n=307) 15.75 $p < 0.001$. On the other hand, when it comes to Time 5 in the Observation condition, slightly more participants select the correct normative answer for Time 4 (82.7% answer 'less') than for Time 5 (76.9 % select 'more'). This difference is in accordance with the hypothesis and it is significant: $\chi^2(1, n=307) = 6.55$, $p = 0.01$.

When it comes to the Intervention condition, a very similar number of participants select the correct answer for Time 2 (69.8% select 'same') and for Time 1 (72% selects 'same'. This difference is not significant: $\chi^2(1, n=321) = 0.724$, p

=0.395. On the other hand, contrary to the hypothesis, a greater number selects the correct answer for Night 5 (77.3% select 'less') than for Night 4 (50.8% select 'less'). This difference is significant: $\chi^2(1, n=321) = 128.1$ $p < 0.001$.

Hypothesis 5: Participants' causal inferences about Time 1 and Time 5 will be conditional and consistent with their inferences about Time 2 and Time 4 respectively.

Causal inferences about Time 1 should be contingent on causal inferences about Time 2. Participants' answers for Time 1 given their answer for Time 2 are displayed in Table 25. As can be seen in Table 25, in accordance to the hypothesis, for the Observation condition, the majority of participants who correctly infer that there would be less tuna at Time 2, also correctly infer that there would more tuna at Time 1 (N=65.15%). For the Intervention condition, the majority of participants who correctly infer that there would be the same number of tuna at Time 2 also correctly infer that there would be the same number of tuna at Time 1 (57.94%).

Similarly, causal inferences about Time 5 should be contingent on causal inferences about Time 4. Participants' answers for Time 5 given their answer for Time 4 are displayed in Table 26. As can be seen in Table 26, in line with the hypothesis, for both conditions, the majority of participants who correctly infer that there would be less tuna at Time 4 also correctly infer there would be more Tuna at 5 (Intervention condition: N = 37.4%; Observation condition: 65.5%).

Table 25.

Table showing percentage of participants' answers to Time 1 as a function of their answer to Time 2. Values highlighted in bold indicate the correct normative answers. Experiment 4, negative loop scenario.

Answer for Time 2	Answer for Time 1	Intervention	Observation
Less	Less	3.43% (N=11)	2.61% (N=8)
	More	6.85% (N=22)	**65.15% (N=200)**
	Same	9.03% (N=29)	4.89% (N=15)
More	Less	1.56% (N=5)	2.61% (N=8)
	More	4.36% (N=14)	13.03% (N=40)
	Same	4.98% (N=16)	2.61% (N=8)
Same	Less	5.61% (N=18)	0.65% (N=2)
	More	6.23% (N=20)	4.56% (N=14)
	Same	**57.94% (N=186)**	3.91% (N=12)

Table 26.

Table showing percentage of participants' answers to Time 5 as a function of their answer to Time 4. Values highlighted in bold indicate the correct normative answers. Experiment 4, negative loop scenario.

Answer for Time 4	Answer for Time 5	Intervention	Observation
Less	Less	6.54% (N=21)	9.45% (N=29)
	More	**37.38% (N=120)**	**65.47% (N=201)**
	Same	6.85% (N=22)	7.82% (N=24)
More	Less	1.56% (N=5)	0.65% (N=2)
	More	33.02% (N=106)	4.89% (N=15)
	Same	1.87% (N=6)	1.63% (N=5)
Same	Less	1.25% (N=4)	1.30% (N=4)
	More	6.23% (N=20)	6.19% (N=19)
	Same	4.67% (N=15)	2.61% (N=8)

Hypothesis 6: Participants in Experiment 4 will make more correct causal inferences than participants in the Experiment 3.

Table 27 shows the proportion of correct answers for each time period as a function of experiment. Values highlighted in bold indicate the highest proportion of correct answers between the experiments for each time period. Participants' answers were recoded into two categories: correct and incorrect. These were entered in a chi-square test for independence comparing Experiment 3 with Experiment 4 (results reported in Table 27).

A significantly larger proportion of participants give the correct answer for time periods 2 and 4 in the observation condition in Experiment 4, than in Experiment 3. However, the format of Experiment 4 does not seem to outperform the one of

Experiment 3 for the Intervention condition – in fact, even though not significantly, Experiment 3 yields a higher proportion of correct answers in all other cases.

Table 27.

Percentage of participants who gave a correct answer for the four questions, as a function of experiment. Experiment 3 and 4, negative loop scenario.

Question	Condition	Experiment 3	Experiment 4	Chi-square value	p-value
Time 1	Intervention	**75.20%**	72.00%	0.508	0.48
	Observation	**86%**	82.70%	0.906	0.341
Time 2	Intervention	**71.40%**	69.80%	0.121	0.727
	Observation	59.50%	**72.60%**	8.626	0.003
Time 4	Intervention	**57.90%**	50.80%	2.017	0.155
	Observation	67.20%	**82.70%**	16.79	0.001
Time 5	Intervention	**81.20%**	77.30%	0.867	0.352
	Observation	**81.70%**	76.90%	1.297	0.254

Positive loop scenario

Table 28 shows the percentage of participants (and number) in each answer category for the four questions, as a function of condition. In each table, the 'correct' normative answer for each question is indicated in bold.

Table 28.

Percentage of participants (and number) in each answer category for the four questions, as a function of each condition. Values highlighted in bold indicate the correct normative answers for each condition. Experiment 4, positive loop scenario.

Question	Answer	Intervention (N=321)	Observation (N=307)
Time 1	Less	8.4% (N=27)	9.1% (N=28)
	More	11.5% (N=37)	**79.8% (N=245)**
	Same	**80.1% (N=257)**	11.1% (N=34)
Time 2	Less	13.1% (N=42)	14% (N=43)
	More	10.9% (N=35)	**70.7% (N=217)**
	Same	**76% (N=244)**	15.3% (N=47)
Time 4	Less	19.9% (N=64)	9.4% (N=29)
	More	**69.5% (N=223)**	***73.6% (N=226)***
	Same	*10.6% (N=34)*	16.9% (N=52)
Time 5	Less	6.5% (N=21)	8.8% (N=27)
	More	**86.9% (N=279)**	**81.4% (N=250)**
	Same	*6.5% (N=21)*	*9.8% (N=30)*

The analyses will be discussed in respect to each of the six experimental hypotheses.

Hypothesis 1: Participants will backtrack (change causal inferences according to the counterfactual supposition for both Time 2 and Time 1) only in the Observation condition (and not in the Intervention condition).

As can be observed in Table 28, most participants in the intervention condition do not backtrack when asked about Time 2 (76%) and Time 1 (80.1%) - they state the number of tuna would not have changed according to the counterfactual supposition. On the other hand, in the Standard observation condition, only a few answer ‘same’

when asked about Time 2 (15.3%) and Time 1 (11.1%). A chi-square for independence was run to compare the observed frequency of cases in each of these two conditions for both Time 2 and Time 1. As predicted by the hypothesis, there was a significant association between condition and answer choice for Time 2: χ^2 (2, n=628) = 264.6 $p < 0.001$, phi = 0.65; and for Time 1: χ^2 (2, n=628) = 324.18 $p < 0.001$, phi = 0.72.

Hypothesis 2: Participants will make the same forward inferences in both conditions (change causal inferences according to the counterfactual supposition for both Time 4 and Time 5).

As can be observed in Table 28, in the Intervention condition only a few participants state the number of tuna would be the same at Time 4 (10.6%) and Time 5 (6.5%). A similar pattern of results can be observed in the Observation condition for Time 4 (16.9%) and Time 5 (9.8%). A chi-square for independence was run to compare the observed frequency of cases in these two conditions for both Time 4 and Time 5. Contrary to the hypothesis, there was a significant association between condition and answer choice for Time 4: χ^2 (2, n=628) 16.66, $p < 0.001$, phi = 0.163. On the other hand, in line with the hypothesis, there was no significant association between condition and answer choice for Time 5: χ^2 (2, n=628) 3.62, $p = 0.164$, phi = 0.076.

Hypothesis 3: The answers selected by participants will reflect the correct qualitative causal inference according to the counterfactual supposition.

As can be observed in Table 28, the correct predicted normative answer is indicated in bold and the majority of participants select that answer for each time

period and for all conditions. In every case, the deviation from 0.3 was highly significant (binomial test, $p<0.001$).

Hypothesis 4: Fewer participants will extend their causal inferences to Time 1 and Time 5.

When it comes to the Intervention condition, contrary to the experimental hypothesis, significantly more participants select the correct answer for Time 1 (80.1% select 'same') than for Time 2 (76% selects 'same'): χ^2= (1, n=321) 2.887, p = 0.089. Likewise, for Time 5 in the Intervention condition, a remarkably greater number of participants select the correct answer for Time 5 (86.9 % select 'more') than for Time 4 (69.5% select 'more'): χ^2= (1, n=321) 46.063 $p<0.0001$.

Very similarly, when it comes to the Observation condition, contrary to the experimental hypothesis, significantly more participants select the correct answer for Time 1 (79.8% select 'same') than for Time 2 (70.7% selects 'same'): χ^2= (1, n=321) 12.324, $p < 0.001$. Likewise, for Time 5 in the Intervention condition, a significantly greater number of participants select the correct answer for Time 5 (81.4 % select 'more') than for Time 4 (73.6% select 'more'): χ^2= (1, n=321) 010.38 p =0.00123.

Hypothesis 5: Participants' causal inferences about Time 1 and Time 5 will be conditional and consistent with their inferences about Time 2 and Time 4 respectively.

Table 29 shows participants' answers for Time 1 given their answer for Time 2. As can be seen in Table 29, for the Observation condition, the majority of participants who correctly infer that there would be more tuna at Time 2 also infer that there would be more tuna at Time 1 (83.3%). For the Intervention condition, the

majority of participants who correctly infer there would be the same number of tuna at Time 2 also infer that there would be the same number of tuna at Time 1 (69.2%).

Table 29.

Table showing percentage of participants' answers to Time 1 as a function of their answer to Time 2. Values highlighted in bold indicate the correct normative answers for each condition. Experiment 4, positive loop scenario.

Answer for Time 2	Answer for Time 1	Intervention	Observation
Less	Less	2.8% (N= 9)	5.2% (N= 16)
	More	2.2% (N= 7)	6.2% (N= 19)
	Same	8.1% (N= 26)	2.6% (N= 8)
More	Less	2.2% (N= 7)	1.0% (N= 3)
	More	5.9% (N= 19)	**65.5% (N= 201)**
	Same	2.8% (N= 9)	4.2% (N= 13)
Same	Less	3.4% (N= 11)	2.9% (N= 9)
	More	3.4% (N= 11)	8.1% (N= 25)
	Same	**69.2% (N= 222)**	4.2% (N= 13)

Participants' answers for Time 5 given their answer for Time 4 are displayed in Table 30. As can be seen in Table 30, for both conditions, the majority of participants who correctly infer that there would be more tuna at Time 4 infer that there would be more tuna at Time 5 (Intervention condition: N = 65.4%; Observation condition: 64.5%).

Table 30.

Table showing percentage of participants' answers to Time 5 as a function of their answer to Time 4. Values highlighted in bold indicate the correct normative answers for each condition.

Answer for Time 4	Answer for Time 5	Intervention	Observation
Less	Less	3.7% (N=12)	2.6% (N=8)
	More	14.6% (N=47)	5.9% (N=18)
	Same	1.6% (N=5)	4.2% (N=13)
More	Less	1.6% (N=5)	3.9% (N=12)
	More	**65.4% (N=210)**	**64.5% (N=198)**
	Same	2.5% (N=8)	5.2% (N=16)
Same	Less	1.2% (N=4)	2.3% (N=7)
	More	6.9% (N=22)	11.1% (N=34)
	Same	2.5% (N=8)	3.6% (N=11)

Hypothesis 6: Participants in Experiment 4 will make more correct causal inferences than participants in the Experiment 3.

Table 31 shows the proportion of correct answers for each time period as a function of experiment. Values highlighted in bold indicate the highest proportion of correct answers between the experiments for each time period. Participants' answers were recoded into two categories: correct and incorrect. These were entered in a chi-square test for independence comparing Experiment 3 with Experiment 4 (results reported in Table 31).

Contrary to the experimental hypothesis, a significantly larger proportion of participants give the correct answer for time periods 2 and 4 in the Observation

condition in Experiment 3, than in Experiment 4. The same is true for time period 4 in the Intervention condition.

Table 31.

Percentage of participants who gave a correct answer for the four questions, as a function of experiment. Experiment 3 and 4, positive loop scenario.

Question	Condition	Experiment 3	Experiment 4	Chi-square value	p-value
Time 1	Intervention	75.20%	**80.10%**	1.506	0.219
	Observation	80.2%	79.80%	0.01	0.92
Time 2	Intervention	74.40%	**76.00%**	0.14	0.707
	Observation	**80.90%**	70.70%	5.092	0.024
Time 4	Intervention	**86.50%**	69.50%	13.63	0.001
	Observation	**82.40%**	73.60%	3.986	0.046
Time 5	Intervention	85.00%	**86.90%**	0.317	0.573
	Observation	**87.80%**	81.40%	2.705	0.1

4.17 Discussion

Like the three preceding experiments, Experiment 4 also provided support for the hypothesis that people are able to reason with simple causal loops and make sensible causal inferences accordingly. Experiment 4 aimed to further clarify the counterfactual supposition from Experiment 3 by adding a specification to the

counterfactual supposition: “Therefore suppose that there has been fishing throughout month 3”.

Negative loop scenario

As in the previous experiments, participants had no trouble with level 1 causal inference. They backtracked only in the Observation conditions and not in the Intervention conditions (hypothesis 1). The difference in backtracking between conditions was significant. Secondly, they made similar forward inferences in both conditions for both scenarios. This was true for most time periods – at time 4 there was a significant association between condition and answer choice - the number of participants who state there would be less tuna is significantly greater in the Observation condition (82.7%) than in the Intervention condition (50.8%); many state there would be more tuna (37.1%).

Furthermore, the majority of participants displayed level 2 causal inferences. They made correct causal inferences for each time period and condition (hypothesis 3). These differences are significantly greater than chance for all time periods except for Time 4 in the Intervention condition: only 50.8% of participants select the correct answer (‘less’), which results in a non-significant binomial test.

Part of the explanation as to why fewer participants make a correct causal inference in the Intervention condition can be found in the very nature of the Intervention manipulation. There could be the feeling that an Intervention is open to noise, or errors. In particular, in Experiment 4 the intervention consisted in introducing new tunas into the ocean – potentially, some participants might think that these new tunas might fail to survive in the new environment, or reproduce at a normal rate. Such assumptions would mean part of the tuna would die and therefore

the quota might not be high enough for fishing to resume, resulting in more tuna at Time 4.

Like in Experiment 3, contrary to the experimental hypothesis, overall participants gave more correct answers for time periods 1 and 5 than for time periods 2 and 4. The only exception where the number of correct answers was greater for the closer time period was in the Intervention condition: slightly more participants selected the correct normative answer for Time 4 rather for than for Time 5.

Finally, in order to display level 3 causal reasoning, participants had to make inferences that were coherent with their representation of the causal loop (hypothesis 5). This was indeed the case - conditional analyses showed that the majority of participants' causal inferences about Time 1 and Time 5 were conditional and consistent with their inferences about Time 2 and Time 4 respectively (this was true for both conditions).

To analyze whether the added clarification to the supposition significantly improved reasoning, the results from Experiment 4 were compared directly with those from the Parallel conditions in Experiment 3. Participants gave significantly more correct answers for Time 2 and Time 4 in the Observation condition. This finding supports the idea that at least some of the noise encountered in Experiment 3 was due to confusion about the state of the 'fishing' variable.

Positive loop scenario

As in Experiment 3, all results were in line with all the experimental hypotheses. The results from Experiment 4 were compared directly with those from the Parallel conditions in Experiment 3. Surprisingly, participants gave significantly

more correct answers in Experiment 3 for both Time 2 in the observation condition and for Time 4 in both the Observation condition and the Intervention condition.

4.18 General Discussion

Summary of findings

The aim of the study was to investigate the boundary conditions for human reasoning about causal loops. Given the complex nature of this question, the present research aimed to provide a stepping-stone for further progressive explorations. Therefore the current study started by adopting a bottom-up approach to investigate if people can engage in basic forms of reasoning - proper causal inferences - based on simple representations of causal loops.

In four experiments participants were presented with a simple scenario based on a causal loop. The nature of the loop was either a negative (stabilizing) or a positive (reinforcing) causal loop. Participants were provided with information about the state of the causal factors in the scenario at five different time periods. Following the presentation of these values, participants were presented with either a counterfactual observation or a counterfactual intervention affecting the mid-time period. In both conditions they were then asked to make causal inferences about the two past time periods and about the two future time periods. The causal inferences consisted in estimating how the values of the causal factors might or might not have changed according to the counterfactual supposition. Importantly, the answers required a causal inference in a qualitative format.

The experimental hypothesis was that people are able to reason with simple

causal loops and make sensible causal inferences accordingly. The extent to which people make sensible causal inferences when reasoning with causal loops was formalized into three levels. The first level is simply differentiation between a true causal inference and a mere estimate of covariation. The second level builds on level 1 in that it consists in providing the correct normative inference (aside from the type of inference). Finally, the third level of causal reasoning requires consistency within the model – causal inferences that are coherent with each other.

The results of each experiment, in relation to these three levels are summarized in Table 31. The results are discussed in more detail following each experiment (see Discussions). The fulfillment of each level was contingent on results supporting the experimental hypothesis encompassed within each level. Level 1 encompassed the following two hypotheses: i) participants will backtrack only in the observation condition and not in the intervention condition; and ii) participants will make the same forward inferences in both conditions. Level 2 hypothesized the answers selected by participants will reflect the correct qualitative causal inference according to the counterfactual supposition. Level 3 predicted participants' causal

inferences about Time 1 and Time 5 will be conditional and consistent with their inferences about Time 2 and Time 4 respectively.

Overall people seem to be able to make proper causal inferences based on simple representations of causal loops. The extent to which they manage to do this seems to vary according to the complexity of the loop at hand. Experiment 1 investigated a very basic loop where the factor being manipulated was also the one participants had to base their causal inferences on. Moreover the loop was founded on a scenario most people were probably familiar with (sleep patterns). Accordingly,

participants displayed up to level 3 causal reasoning: correct and consistent causal inferences throughout all time periods.

Experiment 2 was more complicated in that it involved an additional factor – the factor being manipulated was different from the one on which the participants had to base their causal inferences. Additionally, the loop was constructed on a scenario most people were unfamiliar with (predator-prey relations). Accordingly, participants did not display such an advanced form of causal reasoning as in Experiment 1. In particular the Observation condition presented the most difficulties. These were addressed in Experiment 3. Indeed the clarifications that were introduced improved the causal inferences. In fact, in the clarified conditions (the Parallel conditions), participants displayed up to level 3 causal inferences for both negative and positive scenarios. Experiment 4 aimed to clarify the scenario even further. Even though participants displayed level 1 causal inferences, they encountered some difficulty with level 2 (see discussion following the experiment for more details).

Table 31.

Summary of results Experiments 1-4 in relation to the each level of causal reasoning. 'Yes' is used to state that all hypotheses related to the attainment of that level have been satisfied. The notes in each box describe results not in line with the hypotheses.

Exp.	Scenario	Loop	Level 1	Level 2	Level 3
1	Sleeping patterns	Negative	Yes	Yes	Yes
		Positive	Yes	Yes? • Observation condition, Time 1: majority selects the correct answer but not significantly more than chance. Too many participants do not backtrack - select 'same' (41.3%) instead of 'more' (49.3%).	Yes
2	Predator-prey relation	Negative	Yes	No • Observation condition, time 1: majority does not backtrack - selects 'same' (40%) rather than 'less' (34%). • Observation condition, time 5: majority selects 'less' (44%) rather than 'more' (36%). • Intervention condition, time 5: majority selects 'less' (40%) rather than 'more' (38.8%).	No • Observation condition, time 1 conditional on time 2: only a minority of participants who correctly infer that there would be more tuna at time 2 also correctly infer that there would less tuna at time 1 (N=20%) – a similar numbers answers 'same' (22%) and 'more' (18%).

3	Overfishing	Standard Negative	No Time 5: there was a significant association between condition and answer choice - the number of participants who state there would be more tuna is significantly greater in the Intervention condition (88.6%) than in the Observation condition (66.9%).	No • Observation condition, time 2: less than third of participants makes the correct choice. • Observation condition, time 4 - majority selects the right answer, but this is not significantly better than chance. Too many participants select 'more' (38.2%) instead of 'less' (45.6%).	No • Observation condition: the majority of participants infer (wrongly) that there would be more tuna at Time 2 and more tuna at Time 1 (27.2%). This means that fewer participants give consistent correct normative answers; i.e. indicate there would less tuna at Time 2 and then more tuna at Time 1 (16.2%).
		Standard Positive	Yes	Yes	Yes	
		Parallel Negative	Yes	Yes	Yes	
		Parallel positive	Yes	Yes	Yes	
4	Overfishing	Negative	Time 4: there was a significant association between condition and answer choice - the number of participants who state there would be less tuna is significantly greater in the Observation condition (82.7%) than in the Intervention condition (50.8%); many state there would be more tuna (37.1%).	Yes? Intervention condition, Time 4 – majority selects the correct answer but not significantly more than chance. Too many participants select 'more' (37.1%) instead of 'less' (50.8%).	Yes	
		Positive	Time 4 - there was a significant association between condition and answer choice. The number of participants who state there would be more tuna is significantly greater in the Observation condition (73.6%) than in the Intervention condition (69.5%).	Yes	Yes	

Discussion of results

There are at least two potential non – exclusive explanations as to why increased loop complexity may result in greater difficulty with causal inferences. The first one resides in the limits of working memory. The second one is related to the representation of the time lag between cause and effect.

Working memory. Naturally, the more factors a person has to represent, the higher the cognitive load. This is likely to result in noisier reasoning patterns. This explanation has theoretical implications for the mental model paradigm. This sustains that people will seek to minimize what is being represented explicitly in order to minimize working memory load. This does not seem to be happening in the current experiments. It was argued that if this were the case then there would be some evidence for dissipation – the finding that the causal effect is judged to be diminishing as a function of causal links between the point of 'change' and the target effect. The reason why participants were questioned about Time period 1 and Time period 5, was exactly to explore how people might or might not extend the backward and forward causal inferences to time periods further away from the time period affected by the counterfactual. The idea was that if fewer participants extend their causal inferences further than one time point in the past and in the future it might be because they are not representing those further causal relations explicitly. However, the current experiments do not provide any supportive evidence for this assertion. In other words, participants gave similar number of correct inferences for the further time periods as they did for the closer ones. This suggests all links are being represented, hence incurring a cost for working memory. It may very well be that with more factors people may then try and satisfice by reducing what is being represented explicitly. That is certainly an important question for further research, but as things stand at

present the current findings pose another challenge for the theory of mental models.

One feature that might have facilitated explicit representation of causal relations is that the current study required participants to express answers in terms of qualitative judgments. The answers required a qualitative causal inference (as opposed to an exact numerical estimate) in accordance with the idea that people's spontaneous representation of causal relations might be qualitative (Lagnado, 2011; Pearl, 2000). This feature of the study sets it apart from most causal reasoning studies that require quantitative estimates. The qualitative component could assist judgments by simplifying representations and hence their ease of access as well as reasoning.

Time lag. Another explanation as to why increased loop complexity may result in greater difficulty with causal inferences may be that with more complex loops people have trouble representing the timing between the change and the effect. The system dynamic literature has already established that people misrepresent the time lag between cause and effect in a system (e.g. Sterman, 2006). Specifically, the problem seems to be that there is a time delay in feedback processes that results in an 'effect impatience' problem. Typically, in system dynamic studies people are asked to take a control action on a variable to regulate a system (e.g. beer distribution game). Generally they tend to continue to intervene to correct apparent discrepancies between the desired and actual state of the variable even after sufficient corrective actions have been taken to restore equilibrium. The result is overshooting and oscillation.

The problem of time representation that has emerged in the current study is of a slightly different nature than the 'effect impatience' issue. Experiment 3 and 4 indicated that some participants were confused about when the factor affected by the counterfactual supposition (number of tuna) would in turn affect the second factor (fishing). However, this was not due to a delay in effect but simply confusion about

when the change would take place. This brings out an interesting perspective – perhaps the challenge in causal loop representation lies precisely in the representation of the time lag. This assertion could also explain why people perform considerably better with positive causal loops. With positive loops the direction of change is the same (either increasing or decreasing) and it unfolds in a linear fashion (rather than in a sinusoidal fashion as with negative loops). Hence, time frame does not matter in the same way.

It is also likely that part of the difficulty is induced by artificial experimental scenarios such as the one used in the current study (but also in most system dynamic studies). Asking participants to simulate cause and effect involves asking them to represent time. Even though they might represent the pattern of change correctly (i.e. have the correct causal model), the time scale across which such model unfolds might be incorrect. Asking them to report the state of the factors at specific time points means tapping into individual pieces of data within their model and then deducing their model based on those data points. This could result in underestimating the participants' understanding of loops.

Complexity of representation. Lastly, interpretation and discussion of the current results is somewhat limited by the fact that the participants' representations of the mechanisms generating and governing loops were not investigated directly. The scenarios utilized in the current experiments described the causal relations between factors in terms of one causal link. In other words, participants were told that one factor increased or decreased another. What was not described was the mechanism of change underlying the causal link. For example, in the first scenario, they were told that being rested during the day led to less sleep the following night. They were not told, however, how this actually happens. They were not told how rest affects

circadian rhythms and melatonin levels. This means that, even though participants' inferences were correct, they might have been based on different representations of the mechanisms of change. This implies that simplistic representations of these mechanisms (such as in the current experiments) might suffice in some settings but not in others. For instance, being rested leads to less sleep, but only if the person in question has normal bodily functions, is not affected by excessive mental fatigue, does not consume alcohol and so on.

Naturally, the experiments assumed 'normal' conditions and participants responded accordingly. However, in the real world, where normality is relative, people may need to represent the causal relations generating and governing causal loops in greater depth. This would allow more robust and accurate inferences.

Theoretical implications

At present, theoretical accounts of reasoning with causal loops are fragmented. The field of causal reasoning proposes formal accounts of counterfactual thinking with linear structures but not cyclical structures. The domain of system dynamics puts forwards accounts of reasoning with feedback loops but does not provide a theoretical model of causality within this framework. The theoretical implications of the current findings will be discussed in respects to these two domains.

Causal reasoning. These data show that most people obey a rational rule of counterfactual inference, the undoing principle. Sloman and Lagnado (2005) have already shown that when reasoning about the consequences of a counterfactual supposition, most people do not change their beliefs about the state of the normal causes of the event. Instead, they reason as if the mentally changed event is disconnected and therefore not diagnostic of its causes. The experiments by Sloman

and Lagnado (2005) are the only psychological studies that contrast counterfactual observations and counterfactual interventions. The current study extends their findings based on linear causal structures to cyclical causal structures. This has important theoretical implications for models of causal and counterfactual reasoning. The current study extends their findings based on linear causal structures to cyclical causal structures. This has important theoretical implications for models of causal and counterfactual reasoning. However, once again, it must be pointed out that the set of observation conditions employed in the current experiments are not a direct a reflection of the kind of observation conditions formalized in the experiments by Sloman and Lagnado (2005). The observation conditions in the current experiments (especially Experiment 1 and Experiment 2) are characterized by an absence of information about the reason for the change. However, Experiment 3 has addressed this concern by introducing an additional condition ('Parallel' condition), where counterfactual suppositions did not affect the causal loop actually described in the scenario but, affected another loop with the same properties - essentially identical to the one described in the scenario. The idea was that this would avoid confusion about the source of the change implied by the counterfactual observation. Given this Parallel condition yielded results indicating greater clarity on the participants' part (compared to the 'Standard' condition), the same approach was adopted in Experiment 4.

Lucas and Kemp (2012) provide a model of counterfactual reasoning that extends Pearl's formal account (Pearl, 2000). His model of counterfactual reasoning, allows for the presence of a formal operator that enforces the undoing principle. This operator makes it possible to construct representations that afford valid causal induction of causal relations that support manipulation and control. In turn it affords inference about the effect of such manipulation, be it from actual physical

intervention or merely counterfactual thought about intervention. This formal account of causal reasoning has been highly influential but suffers from two limitations as an account of counterfactual reasoning: it does not distinguish between counterfactual observations and counterfactual interventions, and it does not accommodate backtracking counterfactuals (Lucas & Kemp, 2012).

Hence, Lucas and Kemp (2012) presented an extension of Pearl's account that overcomes both limitations. Lucas and Kemp's approach works with causal systems that are represented using functional causal models and allows these systems to be modified via counterfactual interventions. In addition, however, their approach permits a second kind of modification where exogenous variables are altered not because of a counterfactual intervention, but simply because the counterfactual world might have turned out differently from the real world. An important consequence of this difference is that Lucas and Kemps' model alone accounts for backtracking counterfactuals.

The results of the current study warrant the need to extend Lucas and Kemp's model of counterfactual reasoning to account for backtracking counterfactuals with causal loops. Such extension should also address if the conditions under which a generic counterfactual premise is interpreted as an observation or an intervention are different according to whether the underlying causal structure is chain or a loop.

System dynamics. The current data has important theoretical implications for the domain of system dynamics. As discussed in the Introduction, the general stand emerging from the literature is that the observed dysfunction people seem to display in dynamically complex settings arises from systematic 'misperceptions of feedback'. These are argued to result from mental constructs and processes that are dynamically deficient.

That being said the data from the current study clearly shows people can reason with simple dynamic causal structures. Apart from a reduced complexity, the main difference between the current study and previous work in the system dynamic field is that the causal loop used in the current study was made explicit through the scenario. In other words, participants did not have to detect the loop, instead, they were told about it clearly and explicitly with data points portraying its pattern through time.

This is not necessarily the case with some previous studies (e.g. Sterman, 1989a, 1989b; Osman, 2008) where participants had to work out, or learn, the presence of a loop in the system (for example by manipulating the system). Therefore, one alternative possibility to the view that people are linear thinkers is that people may have trouble inferring feedback structures within complex systems, but not necessarily reasoning with them once these structures are detected and represented.

Practical implications

As argued in the Introduction, dynamic systems with causal loops are ubiquitous. Consequently, establishing that people can understand and reason with simple causal loops has several important practical implications for countless domains. These include how to best communicate risks based on vicious cycles (e.g. positive loops related to climate change or personal health), how to ameliorate reasoning in domains that require controlling dynamic systems (e.g. in management) and how to best teach cyclical structure in educational endeavors (e.g. ecology in school).

Communicating. Perhaps the most urgent application of the current findings is based on the idea that effective risk communication is grounded in deep

understanding of the mental models of policy-makers and citizens. The current study implies that risk communication should indeed integrate causal loops. For example, campaigns attempting to mitigate anthropogenic activities that cause climate change should indeed stress the vicious cyclical nature of the problem. Experiments 3 and 4 used a positive loop scenario where overfishing lead to less conservation which in turn led to more overfishing. The overfishing problem is one of the many vicious cycles where people's actions can make a difference. The present results suggest people can understand the destructive cyclical nature of the problem. More importantly, they clearly appreciated the casual component of the problem. This is an encouraging result because it suggests that any sort of communication encompassing a vicious cycle (including health risks or financial risks) should not be wary of really stressing its' dynamic components. Importantly, the present results suggest this should be done mindfully - special attention needs to be paid to ensure time lags between cause and effect are clear. However, in the present study these time lags were not problematic for positive loops, i.e. vicious cycles.

Controlling. In management there is an increasing interest in enhancing decision makers' understanding of the complex and dynamic systems they are required to control. The idea is to develop systems thinking skills. Such skills include understanding how system behavior is generated by causal loops and time delays within the system. Therefore awareness of the limitations of people's causal models of the system is also added to the list of systems thinking skills (Booth-Sweeney & Sterman, 2000). By learning a set of relations that are frequently found in real systems but regularly misinterpreted, people should be better armed to confront these kinds of phenomena in the future. Hence, the current study's findings related to the time-lag problem have some very applied implications. Special attention needs to be paid to

ensure managers and others alike are able to represent the time lag between causes and effects within the system in a correct manner.

Educating. Arguably, the first step towards achieving a more balanced world is through education. School education already attempts to educate people about phenomena based on complex causal loops: ecology, economics, politics are but a few examples. The current study implies that for this purpose one challenge is to find the best format to facilitate the understanding of loops. This means that loops should not be avoided, but framed in a parsimonious and accessible way. The real key might lie in integrating sufficiently detailed information about the mechanisms of change in a simple framework.

Experimental considerations

The extent to which people make sensible causal inferences when reasoning with causal loops was formalized into three levels. This three-level classification system has practical implications in itself. Primarily, it can serve as benchmark for future studies by providing a clear taxonomy for causal inferences. This can be useful experimentally both for comparing results and for facilitating communication.

On the other hand, a limitation of the current study lies in its somewhat artificial laboratory nature. In the context of causal reasoning this is not an exception, as most studies are based on abstract scenarios. However, as the ultimate goal is to apply the insight gained from such research to real-world thinking, the rather unnatural framing is nonetheless a shortcoming of the study.

Future research

There are three main avenues that future research ought to explore. The first

one is certainly building on the current studies to investigate causal loops of higher complexity. This entails namely loops comprising more than two factors. Following from this idea, causal reasoning with loops nested within linear causal structures is also of great interest. Equally, the study of loops within loops might reveal interesting patterns of inference.

The second path for future efforts should consider that the current study demonstrated inferences based only a deterministic causal system. It is of paramount importance to explore how such findings might apply to probabilistic causal relations as well.

Lastly, eventually it will be critical to conduct studies of a more behavioral nature to explore how reasoning with causal loops is related to judgments and decisions. This venture could adopt an approach based on individual differences, linking participants' causal loops and causal inferences to an applied dependent variable. This sort of methodology could fit well in study of system dynamics.

Conclusion

Are some of the world's environmental problems ultimately just different facets of one single crisis, a crisis of causal reasoning? The findings presented in this study suggest that there is hope: people can reason properly with causal loops. Whether people can make sensible decisions based on the systems that nest these loops is another question. However, overall the outlook is optimistic – it seems that the real challenge will become one of *how* rather than *if*. That is how to communicate causal systems and loops in a way that people are able to represent and reason effectively. Such an endeavor is bound to be a driving force for the applied sciences seeking to advance solution–oriented approaches in many domains.

Chapter 5: General discussion

5. 1 Theoretical implications

How do we represent our world and how do we use these representations to reason about it? The three studies reported in the current thesis explored different aspects of the answer to this question. Even though these investigations offered diverse angles, they originated from the same psychological theory of representation and reasoning. The central idea is that people represent the world and reason about it by constructing dynamic qualitative causal networks. The introduction to the thesis began by laying out the main guiding principles of this tenet: i) the network structure of representations, ii) their qualitative nature; and iii) their dynamic quality. The three studies have been shaped by these principles and in turn offer substantial theoretical implications for each of them. These implications have already been discussed in detail following the report of each study. At this stage it is useful to review them in relation to these three principles, from a more generic perspective.

Causal networks

The first two studies investigated explicitly the idea that people's representations take the form of a network. The first study suggested that mock jurors represent the evidence of a criminal case by arranging it into a casual network. The network structure of the representation was inferred from participants' inferences and how these were extended in line with the hypothesized network structure. The second study suggested people spontaneously represent the causes of an environmental problem in a network structure. Both studies proceeded to show that people can make causal inferences based on their network representations.

An important theoretical implication derived from the first study, is that the network structure can be very much ubiquitous in the sense that it does not have to be limited to strictly causal representations. Even though people represent invariance by focusing on causal relations, the components of the network themselves, do not necessarily need to be causes. In other words, as can be seen in the first study, they can be pieces of evidence or hypotheses about guilt. Similarly there could be other hypotheses about evidence or even ideas, goals and motivations. Therefore, they can potentially take any form required for the reasoning task at hand. This idea implies that that the qualitative causal network framework can really be the backbone of cognition in general, rather than being limited just to causal reasoning. In other words, it is possible that causal models serve as a language of thought from which a lot of mental phenomena arise (e.g., Sloman, 2009). (Naturally, the mind also engages in mental processes such as arithmetic, or grammatical language, which are non-causal).

The second study showed people were able to generate sound network diagrams, suggesting the network structure is spontaneous. The finding that counterfactual judgments were most related to the strength of the represented causal relations, as a function of both direct and indirect links, implies people are able, to some degree, to recruit the whole causal network related to the judgment in question. The main theoretical implication from this finding is that when people engage in causal reasoning about a phenomenon, they can and do recruit a whole causal model as opposed to just individual direct causal relations. Therefore, this idea, besides reinforcing the dynamic qualitative causal network account, denotes the importance of studying causal beliefs, or any pattern of inference for that matter, as a function of its broader overall dynamic causal structure.

Qualitative causal relations.

All three studies have suggested that people's representation and reasoning patterns may take a qualitative nature. The second study showed people's efficacy in recruiting entire causal models to form judgments. This task would not be possible if people had used quantitative ideas to estimate causal strength. Even though participants were indeed asked to provide numerical estimates of strength, they would not have been able to take into account all their estimates in a quantitative fashion (of direct and indirect causal relations) when making the counterfactual judgments. In a similar fashion, the third study required participants to make qualitative counterfactual judgments based on causal loops. Taken together, these findings support the idea that the human mind is smart: it knows how to deal with complex data and simplify it to a degree it can process effectively.

Dynamic models

Finally, the qualitative causal networks investigated throughout the three studies all had dynamic components, providing support for the idea that people can represent and reason about a world that is constantly changing. In the first study, mock jurors easily updated their representations in the face of new evidence; they also made causal inferences in line with the updated representation. In the second study people spontaneously included causal loops into their network diagrams, suggesting they appreciate dynamic causal relations involved in complex problems. The third study directly investigated people's ability to reason with such causal loops. Indeed it showed people are able to make proper causal inferences based on simple representations of causal loops. The extent to which they manage to do this seems to vary according to the complexity of the loop at hand.

To some extent, it seems fair to say that current frameworks of reasoning and representation have overlooked the dynamicity of people's models and reasoning patterns. Even the causal model framework is mostly tacit in respect to the presence of feedback loops. However, given people obviously represent dynamically, the main theoretical implication is simply to take this idea into existing frameworks and develop them to account for this component.

5. 2 Practical implications

The current findings are very important for theoretical frameworks but perhaps even more significant is what they mean in terms of practical implications. Each study's applications to real world situations, both domain-specific as well as generic, have been discussed in detail following each report. However, there are at least three practical implications that are common to all three studies and worth discussing further.

Understanding causal models

The three studies have shown over several applied fields, ranging from juror decision-making to environmental problems, that causal models may be the key to unlocking reasoning in any given domain. In other words, in order to understand how people reason and therefore make decisions about a certain phenomenon, it is important to understand their underlying causal model of that phenomenon. This implies that disciplines seeking to understand human judgment in order to learn the forces that shape it, should make causal models an integral part of their approach. Environmental psychology, for example, aims to understand judgment to nudge it

towards sustainable behaviour. The second study for instance, showed which factors people consider relevant to overfishing. This provides stepping-stones for developing research into understanding why they hold these beliefs and how they might be changed. Similarly, health psychology aims to understand how to increase people's choices geared towards wellbeing. These disciplines, and many others, would benefit from making understanding people's causal models one of their primary goals.

Communication

Given that causal models may be the key to fostering behaviour, their understanding also leads to improved communication with the public. The second study showed how uncovering people's beliefs about an environmental problem facilitates communication about it. In particular, it reveals which aspects of the problem need to be stressed or emphasized to convene better understanding. The second study showed that communication about overfishing needs to emphasize the strength of causal relations. The third study revealed that campaigns attempting to mitigate anthropogenic activities that cause climate change, for example, should indeed stress the vicious cyclical nature of the problem.

Education and learning

A third practical implication derived directly from the three studies is their relevance for education and development of related learning tools. The idea that people make sense of the world by representing it via a causal model, suggests that learning must, to some extent at least, occur in this fashion as well. Learning about the world and its systems can be thought of as a process of updating current casual models with new information. Given this process of dynamic updating, it makes sense

to tailor at least some types of education to tap directly into this concept. To some extent, this is already happening in large corporations where managers get taught how to control systems they need to operate via exposure to system dynamics (Graham, Morecroft, Senge and Sterman (1992). This type of learning, where people are shown and explained models of the systems they need to be operating, does not need to be limited to managerial applications. Educating people about complex environmental problems for instance, could involve explaining causal diagrams of the systems underlying the issue. Similarly, patients could be provided with causal models of diseases, linking life choices to wellbeing. This approach would involve developing the suitable tools for this type of education. Naturally, the studies suggest that these should be based around a qualitative causal network. Surely education and learning would be facilitated if it used the same format people spontaneously use to represent information. The second study suggests diagrams may be the way forward, as people are able to generate these easily and reason through them. The third study points out that formatting loops in simple terms may also be effective.

5.3 Experimental considerations

The three studies presented in the current thesis explored the question of representation and reasoning using simple laboratory based methodologies. Even though these were thorough and accomplished the task of providing stepping-stones for further research, they were still based on very basic representations of the world. The first study utilized mock jurors instead of actual jurors. Naturally a juror is a naïve person so one would hope any conclusions drawn on mock jurors would apply to actual jurors, however it is also possible that representation and reasoning might take a different form in real jury settings. Certainly, criminal cases would be more complex and hence there would be more evidence to represent and reason about. In

the same way, environmental problems such as the one utilized in the second study are much more complex in that they involve many more factors. In addition, participants in the second study were given the factors to represent – in everyday life people often come up with these factors themselves. Again this might interfere with how representations are formed and reasoned about. The third study was also based on very simple causal loops and the question of what would happen with more complex loops, at present, remains unanswered. One priority of future research stemming from these studies is to extend them to settings of higher complexity.

5. 4 Future directions

This thesis started with one pervasive question: How do we represent our world and how do we use these representations to reason about it? The investigations that have been carried out to explore the answer to this query have inevitably led to more questions. The most prominent one of these digs deeper into the origin of representations. The three studies addressed the question of how people represent but they did not tackle the question of how these representations are formed in the first place. In other words, how do people decide which information is relevant for their representation? Such a question was beyond the scope of the thesis but, nonetheless, it is bound to have crucial implications for how people represent and reason. The criteria, whether explicit or implicit, that people apply to define the threshold of what piece of information gets represented into their model, are likely to affect how the piece of information gets represented. Specifically, it may alter the degree of causal strength, or the number of causal connections departing from or going to the factor.

There are many questions still to be answered within the causal model framework. The current thesis suggests that qualitative dynamic causal networks stand at their epicenter and provide a solid platform for launching into future endeavors. These endeavors have enormous potential to elucidate the workings of the human mind and merit priority in cognitive science and psychological research in general. Importantly, they are also, at least partly, the key to unlocking understanding of how to foster sustainable and pro-social behaviour that will contribute to achieving a greater balance in many aspects of today's world.

References

Ajzen, I. (1991). The theory of planned behavior. *Organizational behavior and human decision processes*, *50*(2), 179-211.

Amaya, A. (2007). Formal models of coherence and legal epistemology. *Artificial Intelligence and Law*, 15, 429-447.

Atran, Medin, Hagmayer, Y. and de Kwaadsteniet, L. (2008). Creating causal models: How therapists analyze clients' problems. *In Proceedings and abstracts of the 79th Annual Meeting of the Eastern Psychological Association* (p. 25). Piscataway, NJ: EPA.

Atran, S., Medin, D. L., and Ross, N. O. (2005). The cultural mind: environmental decision making and cultural modeling within and across populations. *Psychological review*, *112*(4), 744.

Axelrod, R. (1976). *Structure of Decision*. Princeton, N.J.: Princeton University Press.

Bamberg, S., and Möser, G. (2007). Twenty years after Hines, Hungerford, and Tomera: A new meta-analysis of psycho-social determinants of pro-environmental behaviour. *Journal of environmental psychology*, *27*(1), 14-25.

Belk, R., Painter, J., and Semenik, R. (1981). Preferred solutions to the energy crisis as a function of causal attributions. *Journal of Consumer Research*, 306-312.

Brehmer, B. (1992). Dynamic decision making: Human control of complex systems. *Acta Psychologica*, 81, 211-241.

Brogan, A. and Hevey, D (2010). Network Analysis. In N.J. Salkind (Ed.), *Encyclopedia of Research Design* (pp. 1776). Kansas: SAGE Publications, Inc.

Byrne, M. (1993). The Convergence of explanatory Coherence and the Story Model: A Case Study in Juror Decision. *Proceedings of the Seventeenth Annual Conference of the Cognitive Science Society*, 539-543.

Cuthbert, L., du Boulay, B., Teather, D., Teather, B., Sharples, M., and du Boulay, G. (1999). Expert/novice differences in diagnosic medical cognition - A review of the literature. *Cognitive Science Research Papers CSRP 508,* School of Cognitive & Computing Sciences, University of Sussex.

Dunlap, R., Van Liere, K., Mertig, A., & Jones, R. E. (2000). Measuring endorsement of the New Ecological Paradigm: A revised NEP scale. *Journal of Social Issues, 56,* 425-442.

Fenton, N., Neil, M., and Lagnado, D. A. (2013). A General Structure for Legal Arguments About Evidence Using Bayesian Networks. *Cognitive science*, *37*(1), 61-102.

Funke, J., 1991. *Solving complex problems: Exploration and control of complex systems*. In: Sternberg, R., Frensch, P. (Ed.), Complex Problem Solving: Principles and Mechanisms. Hillsdale, Lawrence Erlbaum Associates, NJ.

Gerstenberg, T., & Lagnado, D. A. (2010). Spreading the blame: The allocation of responsibility amongst multiple agents. *Cognition*, *115*(1), 166-171.

Gilovich, T., Griffin, D., & Kahneman, D. (Eds.). (2002). *Heuristics and biases: The psychology of intuitive judgment*. Cambridge University Press.

Graham, A. K., Morecroft, J. D., Senge, P. M., & Sterman, J. D. (1992). Model-supported case studies for management education. *European Journal of Operational Research*, *59*(1), 151-166.

Green, D. W. (1997). Explaining and envisaging an ecological phenomenon. *British Journal of Psychology*, 88, 199–217.

Green, D. W. (2001). Understanding microworlds. *Quarterly Journal of Experimental Psychology, 54A*, 879–901.

Green, D. W., & McManus, I. C. (1995). Cognitive structural models: The perception of risk and prevention of coronary heart disease. *British Journal of Psychology, 86*, 321–

336.

Green, D. W., Muncer, S. J., Heffernan, T., & McManus, I. C. (2003). Eliciting and representing the causal understanding of a social concept: A methodological and statistical comparison of two methods. *Papers on Social Representations, 12*, 2.1–2.23

Green, D.W., McManus, I. C. and Derrick, B. J. (1998). Cognitive structural models of unemployment and employment. *British Journal of Social Psychology*, 37, 415–438.

Harris, A. J. L., and Hahn, U. (2009). Bayesian rationality in evaluating multiple testimonies: Incorporating the role of coherence. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 35, 1366-1372.

Hilborn, R. 2012. *Overfishing: What Everyone Needs to Know*. Oxford University Press.

Johnson-Laird, P. N., and Byrne, R. M. (1991). *Deduction*. Hove. East Sussex: Lawrence Erlbaum.

Joireman, J. A., Lasane, T. P., Bennett, J., Richards, D., and Solaimani, S. (2001). Integrating social value orientation and the consideration of future consequences within the extended norm activation model of proenvironmental behavior. *British Journal of Social Psychology*, *40*, 133-155.

Joireman, J. A., Van Lange, P. A. M., Van Vugt, M., Wood, A., Vander Leest, T., and Lambert, C. (2001). Structural solutions to social dilemmas: A field study on commuters' willingness to fund improvements in public transit. *Journal of Applied Social Psychology, 31*, 504-526.

Joireman, J., Balliet, D., Sprott, D., Spangenberg, E., and Shultz, J. (2008). Consideration of future consequences, ego-depletion, and self-control: Support for distinguishing between CFC-immediate and CFC-future sub-scales. *Personality and Individual Differences, 48*, 15-21.

Joireman, J., Posey, D., Truelove, H. B., and Parks, C. D. (2009). The environmentalist who cried drought: Reactions to repeated warnings about depleting resources under conditions of uncertainty. *Journal of Environmental Psychology, 29*, 181-192.

Joireman, J., Shaffer, M., Balliet, and Strathman (2012). Promotion orientation explains why future oriented people exercise and eat healthy: Evidence from the two-factor consideration of future consequences 14 scale. *Personality and Social Psychology Bulletin, 38*, 1272-1287.

Joireman, J., Van Lange, P. A. M., and Van Vugt, M. (2004). Who cares about the environmental impact of cars? Those with an eye toward the future. *Environment and Behavior, 36*, 187-206.

Kahneman, D., Slovic, P., and Tversky, A. (Eds.). (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge University Press.

Kempton, W. (1986). Two theories of home heat control. *Cognitive Science, 10*, 75–90.

Kim, N. S. (2005). Stability and instability over time in explanatory theories of concepts. In *Proceedings of the Twenty-Seventh Annual Conference of the Cognitive Science Society* (p. 2500).

Kim, N. S., and Ahn, W. K. (2002a). Clinical psychologists' theory-based representations of mental disorders predict their diagnostic reasoning and memory. *Journal of Experimental Psychology: General*, *131*(4), 451.

Kim, N. S., and Ahn, W. K. (2002b). The influence of naive causal theories on lay concepts of mental illness. *American Journal of Psychology,* 115(1), 33-66.

Kim, N., Luhmann, C.C, Pierce, M.L. and Ryan, M.M., (2009). The conceptual centrality of causal cycles. *Memory & Cognition, 37(6),* 744-758.

Kollmuss, A., and Agyeman, J. (2002). Mind the gap: why do people act environmentally and what are the barriers to pro-environmental behavior? *Environmental education research*, *8*(3), 239-260.

Kortenkamp, K. V. and Moore, C. F. (2006). Time, uncertainty, and individual differences in decisions to cooperate in resource dilemmas. *Personality and Social Psychology Bulletin, 32*, 603-615.

Lagnado, D. A., Fenton, N., and Neil, M. (2013). Legal idioms: a framework for evidential reasoning. *Argument & Computation*, *4*(1), 46-63.

Lagnado, D. A. (2011). *Thinking about Evidence*. In Dawid, P., Twining, W., Vasaliki, M. (ed.) Evidence, Inference and Enquiry. Oxford University Press/British Academy.

Lagnado, D.A., Waldmann, M.R., Hagmayer, Y., and Sloman, S.A. (2007). *Beyond covariation: Cues to causal structure*. In Gopnik, A., & Schultz, L. (eds.), Causal learning: Psychology, Philosophy, and Computation, pp. 154–172. Oxford: Oxford University Press.

Latane, B., and Darley, J. M. (1968). Group inhibition of bystander intervention in emergencies. *Journal of Personality and Social Psychology; Journal of Personality and Social Psychology*, *10*(3), 215.

Lunt, P. K. (1988). The perceived causal structure of examination failure. *British Journal of Social Psychology*, 27, 171–179.

Lewis, D. K. (1986). *On the plurality of worlds* (Vol. 322). Oxford: Blackwell.

Lucas, C. G., and Kemp, C. (2012). A unified theory of counterfactual reasoning. In *Proceedings of the 34th Annual Meeting of the Cognitive Science Society.*

McElwee, R. O. B., and Brittain, L. (2009). Optimism for the world's future versus the personal future: Application to environmental attitudes. *Current Psychology*, *28*(2), 133-145.

Milfont, T. L., and Duckitt, J. (2010). The environmental attitudes inventory: A valid and reliable measure to assess the structure of environmental attitudes. *Journal of Environmental Psychology*, *30*(1), 80-94.

Milfont, T. L., and Gouveia, V. V. (2006). Time perspective and values: An exploratory study of their relations to environmental attitudes. *Journal of Environmental Psychology, 26*, 72-82.

Moxnes, E. (1998). Overexploitation of renewable resources: The role of misperceptions. *Journal of Economic Behavior & Organization*, *37*(1), 107-127.

Moxnes, E. (2000). Not only the tragedy of the commons: misperceptions of feedback and policies for sustainable development. *System Dynamics Review*, *16*(4), 325-348.

Nowack, K., Milfont, T. L., and van der Meer, E. (2012). Future versus Present: Time Perspective and Pupillary Response in a Relatedness Judgment Task investigating temporal event knowledge. *International Journal of Psychophysiology*.

Olsson, E.J. (1998). Making Beliefs Coherent. *Journal of Logic, Language and Information*, 7, 143-163.

Osman, M. (2008). Seeing is as Good as Doing. *The Journal of Problem Solving*, 2,1.

Osman, M. (2010). Controlling uncertainty: a review of human behavior in complex dynamic environments. *Psychological Bulletin*, *136*(1), 65.

Pearl, J. (1988). Embracing Causality in Default Reasoning. *Artificial Intelligence*, 35 (2), 259-271.

Pearl, J. (1998). *On the definition of actual cause*. Technical Report R-259, Computer Science Dept., UCLA.

Pearl, J. (2000). *Causality: Models, Reasoning and Inference.* Cambridge: Cambridge University Press.

Pennington, N. & Hastie, R (1993). Reasoning in Explanation based Decision Making. *Cognition*, 49, 123-163.

Pennington, N. and Hastie, R. (1981). Juror Decision Making Models: The Generalisation Gap. *Psychological Bulletin*, 89, 246-287.

Pennington, N. and Hastie, R. (1986). Evidence Evaluation in complex decision-making. *Journal of Personality &* Social Psychology, 51, 242-258.

Pennington, N. and Hastie, R. (1988). Explanation Based Decision Making: Effects of memory Structure on Judgment. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 14, 521-533.

Pennington, N. and Hastie, R. (1992). Explaining the Evidence: Tests of the Story Model for Juror Decision Making. *Journal of Personality & Social Psychology*, 62 (2), 189-206.

Rehder, B., and Martin, J. B. (2011). A generative model of causal cycles. *In Proceedings of the 33rd annual conference of the cognitive science society.* Austin, TX: Cognitive Science Society.

Rein, J. R., Love, B. C., and Markman, A. B. (2007). Feature relations and feature salience in natural categories. In *Proceedings of the Twenty-Ninth Annual Conference of the Cognitive Science Society* (pp. 593-598.).

Rottman, B. M. and Keil, F. C. (2012). Causal structure learning over time: Observations and Interventions. *Cognitive Psychology*, 64, 93-125.

Rottman, B. and Hastie, R. (2013). Reasoning About Causal Relationships: Inferences on Causal Networks. *Psychological bulletin, 32, 1-23.*

Simon, D. (2004). A Third View of the Black Box: Cognitive Coherence in legal Decision Making. *Memory and Cognition*, 71 (2), 511-586.

Simon, D., and Holyoak, K. J. (2002). Structural dynamics of cognition: From consistency theories to constraint satisfaction. *Personality & Social Psychology Review*, 6, 283-294.

Simon, D., Snow, C., and Read, S. J. (2004). The redux of cognitive consistency theories: Evidence judgments by constraint satisfaction. *Journal of Personality and Social Psychology, 86*, 814–837.

Sloman, S. (2009). *Causal models: How people think about the world and its alternatives*. Oxford University Press.

Sloman, S. A., Love, B. C., and Ahn, W. K. (1998). Feature centrality and conceptual coherence. *Cognitive Science*, *22*(2), 189-228.

Sloman, S.A. and Lagnado, D.A. (2005). Do we 'do'? *Cognitive Science, 29*, 5-39.

Smith VL, Suchanek GL, Williams AW. (1988) Bubbles, Crashes, and Endogenous expectations in experimental spot asset markets. *Econometrica: Journal of the Econometric Society*, 1119-1151.

Spirtes, P., Glymour, C., and Scheines, R. (1993). Causality, prediction, and search. *Lecture Notes in Statistics*, *81*.

Steiner, I. D. (1972). *Group process and productivity*. New York: Academic Press.

Sterman, J. D. (1989). Modeling managerial behavior: Misperceptions of feedback in a dynamic decision making experiment. *Management science*, *35*(3), 321-339.

Sterman, J. D. (2006). Learning from evidence in a complex world. *American journal of public health*, *96*(3), 505-514.

Sterman, J. D., and Sweeney, L. B. (2002). Cloudy skies: assessing public understanding of global warming. *System Dynamics Review*, *18*(2), 207-240.

Stern, P. C., Dietz, T., Abel, T., Guagnano, G. A., and Kalof, L. (1999). A value-belief-norm theory of support for social movements: The case of environmentalism. *Research in*

Human Ecology, 6 (2), 81-97.

Stratham, A., Gleicher, F., Boninger, D. S., and Edwards, C. S. (1994). The consideration of future consequences: Weighing immediate and distant outcomes of behavior. *Journal of Personality and Social Psychology, 66*(4), 742-752.

Sweeney, L. B., and Sterman, J. D. (2000). Bathtub dynamics: initial results of a systems thinking inventory. *System Dynamics Review*, *16*(4), 249-286.

Tangari, H.T. and Smith, R.J. (2012). How the Temporal Framing of Energy Savings Influences Consumer Product Evaluations and Choice. *Psychology and Marketing, 29(4),* 198-208.

Thagard, P. (1989). Explanatory Coherence. *Behavioural Brain Sciences*, 12, 435-467.

Thagard, P. (2000). *Coherence in thought and action*. Cambridge, MA: MIT Press.

Thogersen, J. (2004). A cognitive dissonance interpretation of consistencies and inconsistencies in environmentally responsible behavior. *Journal of Environmental Psychology, 24* (1), 93-103. University Press.

Waldmann, M. R., and Hagmayer, Y. (2005). Seeing versus doing: two modes of accessing causal knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(2), 216.

Waldmann, M. R., Hagmayer, Y, and Blaisdell, A. P. (2006). Beyond the information given: Causal models in learning and reasoning. *Current Directions in Psychological Science, 15 (6),* 307-311.

Waldmann, M. R., Hagmayer, Y., and Blaisdell, A. P. (2006). Beyond the Information Given Causal Models in Learning and Reasoning. *Current Directions in Psychological Science*, *15*(6), 307-311.

Wellman, M. P. and Henrion, M. (1993). Explaining 'Explaining Away'. *IEE Transaction on Pattern Analysis and Machine Intelligence*, 15 (3), 287-292.

White, P. A. (1992). The anthropomorphic machine: Causal order in nature and the world view of common sense. *British Journal of Psychology, 83,* 61–96.

White, P. A. (1995). Common sense construction of causal processes in nature: A causal network analysis. *British Journal of Psychology, 86,* 377–395.

White, P. A. (1997). Naive ecology: Causal judgements about a simple ecosystem. *British Journal of Psychology, 88*, 219–233.

White, P. A. (1997). Naive ecology: Causal judgements about a simple ecosystem. *British Journal of Psychology, 88*, 219–233.

White, P. A. (1998). The dissipation effect: A general tendency in causal judgments about complex physical systems. *The American journal of psychology*, 111(3), 379-410.

White, P. A. (1999). The dissipation effect: A naive model of causal interactions in complex physical systems. American Journal of Psychology, 112*, 331–364.*

White, P. A. (1999). The dissipation effect: A naive model of causal interactions in complex physical systems. *American Journal of Psychology*, 112*, 331–364.*

White, P. A. (2000). Naive analysis of food web dynamics: a study of causal judgement about complex physical systems. *Cognitive Science, 24,* 605-650.

White, P. A. (2008). Beliefs about interactions between factors in the natural environment: a causal network study. *Applied Cognitive Psychology, 22,* 559-572.

White, P. A. (2008). Beliefs about interactions between factors in the natural environment: A causal network study. *Applied Cognitive Psychology*, *22*(4), 559-572.

Wigmore, J. H. (1913). Problem of Proof. *Illinois Law Review*, *8 (2)*, 77-103.

Woodward, J. (2000). Explanation and invariance in the special sciences. *The British Journal for the Philosophy of Science*, *51*(2), 197-254.

Appendix I

Generalized Liner Mixed Model Analysis

Experiment 1

A generalized linear mixed effect model was fitted with binomial errors and a logit link (using SPSS). This was done separately for the negative loops scenario and the positive loop scenario (comparison between the two was not deemed appropriate as the two scenarios differ on too many dimensions). For each scenario, the dependent variable was the probability of a correct response (each participant's response was recoded as being correct or incorrect). The predictors were condition (observation or intervention) and time period question (Time 1, Time 2, Time 4 and Time 5).

Negative loop scenario. The model was significant: $F(7,840) = 44.85$, $p<0.001$. There was a main effect of Time period question: $F(3,840) = 45.82$, $p<0.001$. There was no main effect of Condition: $F(1, 840)= 0.009$ $p=0.925$. There was a significant 2-way interaction between Condition and Time period question: $F(3,840)=39.93$, $p<0.001$. Pairwise contrasts were conducted to tease apart the significant 2-way interaction. There was a significant difference between conditions for Time period 1 ($z=16.763$, $p<0.001$) and Time period 2 ($z=17.3$, $p<0.001$). There was no significant difference at Time period 4 ($z=0.25$, $p=0.8$) and Time period 5 ($z=0.001$,$p=0.12$).

Positive loop scenario. The model was significant: $F(7,840) = 24.41$, $p<0.001$. There was a main effect of Time period question: $F(3,840) = 21.32$, $p<0.001$. There was a main effect of Condition: $F(1, 840)= 48.33$ $p<0.001$. There was a significant 2-way interaction between Condition and Time period question:

F(3,840)=15.52, p<0.001. Pairwise contrasts were conducted to tease apart the significant 2-way interaction. There was a significant difference between conditions for Time period 2 (z=13.13, p<0.001) and Time period 4 (z=2.58, p=0.001). There was no significant difference at Time period 1 (z=1.47, p=0.143) ad Time period 5 (z=1.03, p=0.3).

Experiment 2

A generalized linear mixed effect model was fitted with binomial errors and a logit link (using SPSS). The dependent variable was the probability of a correct response (each participant's response was recoded as being correct or incorrect). The predictors were condition (observation or intervention) and time period question (Time 1, Time 2, Time 4 and Time 5). The model was significant: F (7,388) = 5.95, p<0.001. There was a main effect of Time period question: F (3,388) = 8.26, p<0.001. There was no main effect of Condition: F (1, 120)= 0.005 p=0.947. There was a significant 2-way interaction between Condition and Time period question: F(3,840)=5.95, p=0.001. Pairwise contrasts were conducted to tease apart the significant 2-way interaction. There was a significant difference between conditions for Time period 1 (z=3.99, p<0.001) and Time period 4 (z=2.01 p=0.04). There was no significant difference for Time period 2 (z=1.12, p=0.27) and Time period 5 (z=0.28, p=0.77).

Experiment 3

A generalized linear mixed effect model was fitted with binomial errors and a logit link (using SPSS). This was done separately for each of the 4 scenarios: the standard negative loop scenario, the standard positive loop scenario, the parallel negative loop scenario and the parallel positive loop scenario. For each scenario, the

dependent variable was the probability of a correct response (each participant's response was recoded as being correct or incorrect). The predictors were condition (observation or intervention) and time period question (Time 1, Time 2, Time 4 and Time 5). In addition, the standard conditions were compared with the parallel conditions. Therefore 'Standard/Parallel' was used as an additional predictor variable in the analysis.

Standard Negative loop scenario. The model was significant: $F(7,1064) = 22.82$, $p<0.001$. There was a main effect of Time period question: $F(3,1064) = 34.25$, $p<0.001$. There was no main effect of Condition: $F(1, 307)= 55.3$ $p<0.001$. There was a significant 2-way interaction between Condition and Time period question: $F(3,1064)=11.87$, $p<0.001$. Pairwise contrasts were conducted to tease apart the significant 2-way interaction. There was a significant difference between conditions for Time period 1 ($z=5.92$, $p<0.001$), Time period 2 ($z=8.24$, $p<0.001$) and Time period 5 ($z=4.2$, $p<0.001$). Time period 4 was not significant ($z=0.075$, $p=0.94$).

Standard Positive loop scenario. The model was significant: $F(7,1064) = 5.95$, $p<0.001$. There was a main effect of Time period question: $F(3,1064) = 4.39$, $p=0.004$. There was a main effect of Condition: $F(1, 351)= 13.79$, $p<0.001$. There was a significant 2-way interaction between Condition and Time period question: $F(3,1064)=2.58$, $p=0.05$. Pairwise contrasts were conducted to tease apart the significant 2-way interaction. There was a significant difference between conditions for Time period 1 ($z=4.33$, $p<0.001$) and Time period 2 ($z=1.06$, $p=0.007$). There was no significant difference for Time period 4 ($z=1.43$, $p=0.156$) and Time period 5 ($z=0.89$, $p=0.38$).

Parallel Negative loop scenario. The model was significant: F (7,1048) = 39.22, p<0.001. There was a main effect of Time period question: F (3,1048) = 71.48, p<0.001. There was no main effect of Condition: F (1, 279)= 28.68 p<0.001. There was a significant 2-way interaction between Condition and Time period question: F(3,1048)=14.01, p<0.001. Pairwise contrasts were conducted to tease apart the significant 2-way interaction. There was a significant difference between conditions for Time period 2 (z=9.59, p<0.001), and Time period 4 (z=4.43, p<0.001). There was no significant difference for Time period 4 (z=0.05, p=0.16) and Time period 5 (z=0.031, p=0.37).

Parallel Positive loop scenario. The model was significant: F (7,1048) = 2.25, p<0.001. There was a main effect of Time period question: F (3,1048) = 4.6, p=0.003. There was no main effect of Condition: F (1, 294)= 1.16, p=0.28. There was no significant 2-way interaction between Condition and Time period question: F(3,1048)=0.39, p=0.76.

Standard versus Parallel Negative loops scenarios. The model was significant: F (15,2112) = 29.06, p<0.001. There was a main effect of Standard/Parallel condition F(1, 596) = p=0.029. There was a main effect of Time period question: F (3,2112) = 97.57, p<0.001. There was a main effect of Condition: F (1, 596)= 82.28 p<0.001. There was a significant 2-way interaction between Condition and Time period question: F(3,2112)=13.1, p<0.001. Importantly, there was a significant 3-way interaction between Standard/Parallel condition, Time and Condition: F (3, 2112)= 10.661, p<0.001.

Pairwise contrasts were conducted to tease apart the significant interactions. For the 2-way interaction between Condition and Time, there was a significant

difference at all time periods: Time 1 (z= 0.15, p<0.001), Time 2 (z= 2.69, p=0.007), Time 4 (z= 0.14, p<0.001) and Time 5 (z= 0.109, p=0.002). For the 3-way interaction, the Intervention conditions of the Standard version and Parallel versions were significantly different at Time period 2: z= 2.69, p=0.007. There was no significant difference at Time period 1 (z=0.62, p=0.5), Time period 4 (z=0.33, p=0.5) and Time period 5 (z=1.5, p=0.13). The Observation conditions of the Standard version and Parallel versions were significantly different at all time periods: Time period 1 (z= 5.44, p<0.001), Time period 2 (z= 3.48, p<0.001), Time period 4 (z= 5.11, p<0.001) and Time period 5 (z= 2.54, p=0.01).

Standard versus Negative Positive loop scenarios. The model was significant: F (15,2112) = 3.83, p<0.001. There was no main effect of Standard/Parallel condition F(1, 645) = p=0.109. There was a main effect of Time period question: F (3,2112) = 8.27, p<0.001. There was a main effect of Condition: F (1, 645)= 11.591, p<0.001. There was no significant 2-way interaction between Condition and Time period question: F(3,2112)=1.36, p=0.254. There was no significant 3-way interaction between Standard/Parallel condition, Time and Condition: F (3, 2060)= 2.06, p=0.104.

Experiment 4

A generalized linear mixed effect model was fitted with binomial errors and a logit link (using SPSS). The dependent variable was the probability of a correct response (each participant's response was recoded as being correct or incorrect). The predictors were condition (observation or intervention) and time period question (Time 1, Time 2, Time 4 and Time 5). In addition, the results from Experiment 3 were compared with those from Experiment 4. Therefore 'Experiment' was used as a

additional predictor variable in the analysis.

Negative loop scenario. The model was significant: F (15,2504) = 114.11, $p<0.001$. There was a main effect of Time period question: F (3,2504) = 145.63, $p<0.001$. There was no main effect of Condition: F (1, 2504)= 0.299 $p=0.59$. There was a significant 2-way interaction between Condition and Time period question: $F(3,2504)=123.83$ $p<0.001$. Pairwise contrasts were conducted to tease apart the significant 2-way interaction. There was a significant difference between conditions for Time period 1 ($z=15.2$, $p<0.001$) and Time period 4 ($z=421.59$, $p<0.001$). There was no significant difference for Time period 2 ($z=1.89$, $p=0.06$) and Time period 5 ($z=0.213$, $p=0.83$).

Positive loop scenario. The model was significant: F (7,2504) = 84.97, $p<0.001$. There was a main effect of Time period question: F (3,2504) = 96.56, $p<0.001$. There was a main effect of Condition: F (1, 2504)= 147.43, $p<0.001$. There was a significant 2-way interaction between Condition and Time period question: $F(3,2504)=97.07$, $p<0.001$. Pairwise contrasts were conducted to tease apart the significant 2-way interaction. There was a significant difference between conditions for Time period 2 ($z=0.65$, $p<0.001$) and Time period 4 ($z=0.7$, $p<0.001$). There was no significant difference for Time period 1 ($z=1.2$, $p=0.23$) and Time period ($z=1.45$, $p=0.14$)

Experiment 3 versus Experiment 4 Negative loop scenarios. The model was significant: F (15, 3552) = 72.87, $p<0.001$. There was a main effect of Experiment $F(1, 3552) = 58.42$, $p<0.001$. There was a main effect of Time period question: F (1,3552) = 138.92, $p<0.001$. There was a main effect of Condition: F (1, 3552)= 72.87, $p<0.001$. As hypothesized, there was a significant 2-way interaction between

Condition and Time period question: F(3,3552)=29.02, p<0.001. Importantly, there was a significant 3-way interaction between Experiment, Time and Condition: F (7, 3552)= 50.38, p<0.001.

Pairwise contrasts were conducted to tease apart the significant interactions. For the 2-way interaction between Condition and Time, there was a significant difference at Time period 1 (z= 8.27, p<0.001), Time period 2 (z= 4.29, p<0.001), and Time period 4 (z= 4.34, p<0.001). There was no significant difference for Time period 5 (z=0.25, p=0.81). For the 3-way interaction, there was a significant difference between the Intervention conditions of Experiment 3 and Experiment 4 for Time period 1 (z= 2.45, p=0.014), Time period 2 (z= 15.27, p<0.001) and Time period 4 (z= 3.6, p<0.001). There was no significant difference for Time period 5 (z=1.13, p=0.26). There was also a significant difference between the Observation conditions of Experiment 3 and Experiment 4 for Time period 1 (z= 15.46, p<0.001), Time period 2 (z= 4.112, p<0.001) and Time period 4 (z= 20.28, p<0.001). There was no significant difference for Time period 5 (z=1.1, p=0.27).

Experiment 3 versus Experiment 4 Positive loop scenarios. The model was significant: F (15,3552) = 43.86 p<0.001. There was a main effect of Experiment F(1, 3552) = p<0.001. There was a main effect of Time period question: F (3, 3552) = 37.18, p<0.001. There was a main effect of Condition: F (1, 3552)= 26.9, p<0.001. There was a significant 2-way interaction between Condition and Time period question: F(3,3552)=23.49, p<0.001. Importantly, there was a significant 3-way interaction between Experiment, Time and Condition: F (7, 3552)= 29.24, p<0.001.

Pairwise contrasts were conducted to tease apart the significant interactions.

For the 2-way interaction between Condition and Time, there was a significant difference at Time period 2 ($z= 7.56$, $p<0.001$) and Time period 4 ($z= 6.57$, $p<0.001$). There was no significant difference for Time period 1 ($z=0.8$. $p=0.41$) and Time period 5 ($z=0.8$, $p=0.41$). For the 3-way interaction, there was a significant difference between the Intervention conditions of Experiment 3 and Experiment 4 for Time period 2 ($z= 3.55$, $p<0.001$) and for Time period 4 ($z= 24.16$, $p<0.001$). There was no significant difference for Time period 1 ($z=0.89$, $p=0.37$) and Time period 5 ($z=0.45$, $p=0.66$). There were no significant differences between the Observation conditions of Experiment 3 and Experiment 4: Time period 1 ($z=1.13$, $p=0.2$), Time period 2 ($z=0.23$, $p=0.8$), Time period 4 ($z=0.32$, $p=0.75$) and Time period 5 ($z=1.13$, $p=0.23$).